ACER Research Monograph

66

The Art of Item Writing: Cross-Curricular Testing at ACER 1964 to 2014

Doug McCurry with
John Harding

Australian Council *for* Educational Research

ACER

ACER Research Monograph No. 66

The Art of Item Writing: Cross-curricular Testing at ACER 1964 to 2014

Doug McCurry
with
John Harding

November 2017

Australian Council *for* Educational Research

ISBN 978-1-74286-401-3

Published 2017 by:
Australian Council *for* Educational Research Ltd
19 Prospect Hill Road, CAMBERWELL VICTORIA 3124 AUSTRALIA

Preferred citation:
McCurry, D. (2017). *The Art of Item Writing: Cross-curricular Testing at ACER 1964 to 2014. ACER Research Monograph 66*. Melbourne, Australian Council *for* Educational Research
http://research.acer.edu.au/acer_monographs/8)

Drawing by Alan Dunn; 1966 The New Yorker Magazine, Inc.

'You see, Dad, Professor McLuhan says that the environment that man creates becomes his medium for defining his role in it. The invention of type created linear, or sequential thought, separating thought from action. Now, with TV and folk singing, thought and action are closer and social involvement is greater. We again live in a village. Get it?'

CSSE Question 90

Some of Marshall McLuhan's statements such as 'The electric light is pure information' have been called 'flamboyantly indecipherable'. Which one of the following quotations from the passage best illustrates this quality?

A. 'Today's television child is attuned to up-to-the-minute 'adult' news . . .'
B. 'Today's child is growing up absurd . . .'
C. 'Many of our institutions suppress all the natural direct experience of youth . . .'
D. 'Education must shift from instruction . . .'

When the first set of tests was prepared for the Commonwealth Secondary Scholarships Examination in 1964 the staff concerned thought that they might lead to new procedures in assessment in schools, but had little time to consider that the questions used in the test papers could also be suitable material for use by teachers in activities other than assessment.

As time has gone on, more and more teachers have commented on the nature of the material used in the tests, and more and more reports have come to the Australian Council for Educational Research about the interest shown by examinees in a good deal of the material. It has become clear that when they are able to examine at leisure the material in the tests devised for the CSSE, many teachers find there is valuable and interesting material to use in their classrooms to encourage, for example, careful critical thinking, to spur the interest of students in the search for and interpretation of evidence, and to draw attention to features of mood or mind that transcend common subject barriers.

W.C. Radford
Director, Australian Council *for* Educational Research, 1972

There were two other outcomes of this experience of constructing test items that were important. The first related to the discourse, the arguments about the best answer that characterised the panel meetings. The second related to the values and effects of this particular testing program, and how to deal with that (Wilson, 1970).

As we got better at writing 'distractors' for multiple choice questions, we found advocates among the 'expert' panel for some of the distractors as the best answer, rather than the one chosen by the test writer. Of more potential educational significance was the argumentation itself, and its effect on our ability to think sharply and clearly within the fields being discussed. Tests themselves can never produce improvement in individual performance; but our experience suggested that argumentative discourse about test items could. A serendipitous piece of research at one school confirmed this. One hundred students thus engaged for about twenty hours raised test scores on each of three multiple choice papers by half a standard deviation, despite the ACER publications that claimed these tests could not be 'taught' (Wilson, 1969).

The second experience related to educational values, and our attempts as 'examiners' to grapple with this. None of the full-time test constructors approved of the Commonwealth Secondary Scholarship tests as an educational intervention. They were a politically inspired election gimmick. We were aware that they would have an influence on what schools taught, and possibly how they taught, even though they were supposed to be 'curriculum free' as well as value free. As a result we took 'educational value' as a major criteria for test validity, at least at the level of our own personal discourse. The material we chose for tests must face the question 'would education be improved if teachers did try to prepare students for this sort of exercise, for answering these sorts of questions on these sorts of information or issues, for engaging in this sort of thinking and problem solving?' I was learning that no test was value free, and that these tests were certainly informed by a (possibly idiosyncratic) view of educational relevance.

Noel Wilson
(Science and mathematics item writer 1964 to 1970) in
Educational Standards and the Problem of Error, 1998

Foreword

The Baby Boomers' Scholarship Test

This monograph is about the development of a distinctive style of cross-curricular testing that was developed at the Australian Council *for* Educational Research (ACER) in the 1960s and 1970s, and is still in use in some testing programs presently.

The impetus for this kind of test development was the decision of the Menzies government to award 10,000 scholarships each year for students to study at Years 11 and 12 through a test called *The Commonwealth Secondary Scholarship Examination* (CSSE). The testing program ran for 10 years, during which around one million candidates sat the test.

This text attempts to demonstrate the view that the CSSE was a distinctively innovative kind of testing. It contrasted markedly with the paradigm of psychometric testing that was preeminent at that time, and continues to be a shaping force in generic skills testing. Many of the CSSE test items manifest a distinctive aim in generic skills testing, and many of the items required genuinely sophisticated, higher-order thinking from candidates.

Through this monograph, it is worth remembering and preserving the CSSE for its importance to ACER and for the new players in the test development field, but it is also hoped that this work will be of interest to those who took the test in the 1960s and 1970s. As part of the preservation process, all of the CSSE test forms are available on the ACER website at http://research.acer.edu.au/csse/ so that those who took the test can see what, if anything, they recall of the experience.

Contents

Exhibits

Tables

Figures

Terminology

When one first comes to work at ACER one experiences acronym shock because programs and concepts are commonly reduced in discussion and writing to acronyms. The following is offered for the assistance of new players.

ACER	Australian Council *for* Education Research
ASAT	Australian Scholastic Aptitude Test
AST	ACT Scaling Test
BSSS	Board of Senior Secondary Studies of the Australian Capital Territory
CEEB	College Entrance Examination Board in the United States
CSSE	Commonwealth Secondary Scholarship Examination
CSSE Humanities	CSSE Comprehension and Interpretation (Humanities)
CSSE Mathematics	CSSE Comprehension and Interpretation (Mathematics)
CSSE Science	CSSE Comprehension and Interpretation (Sciences)
CSSE Writing	CSSE Written Expression
CSTP	Cooperative Scholarship Testing Program
GAMSAT	Graduate Australian Medical School Admissions Test
GAT	General Achievement Test
HASS	Humanities, arts and social sciences
QCST	Queensland Core Skills Test
MST	Mathematics, science and technology
SAT	Previously called The Scholastic Aptitude Test produced for the CEEB
STAT	Special Tertiary Admission Test
TEEP	Tertiary Entrance Examination Project
VBOS	Victorian Board of Studies
VCE	Victorian Certificate of Education

Timeline

This monograph aims to provide a comprehensive account of the 'ACER-type' of generic, cognitive skills testing that developed in the 1960s, and is still evolving in the second decade of the 21st century, with the view of preserving some of the history for future test writers. A summary of the key milestones is captured in the timeline below.

	START	END
Scholastic Aptitude Test (SAT) established in the United States	1926	ongoing
ACER established	1930	ongoing
ACER produced the first non-verbal test of general ability for Australia	1936	1937
Intelligence tests (Grade 6) for NSW Department of Education	1937	1948
Construction of intelligence tests for NSW Public Service Board	1940	1954
Examination of Melbourne Uni students for admission under adult matriculation scheme	1940	1955
Testing of applicants for entry to Melbourne Technical College	1942	1945
Examination for Victorian Junior Scholarships awarded for secondary school courses	1945	1960
General ability tests for Victorian Government scholarship examinations	1945	1963
ACER Junior A test (Grade 4) for NSW Department of Education	1946	1947
ACER intermediate D test	1946	1951
ACER Junior B (verbal) general ability tests	1947	1951
ACER Jenkins non-verbal test: modification and standardization	1948	1953
ACER mechanical reasoning test (13-14 years)	1950	1951
Testing of teachers college entrants	1955	1959
The general ability test for children aged 7 to 9	1959	1961
The general ability tests for the NSW Department of Labour and National Service	1962	1966
Cooperative Scholarship Testing Program (CSTP)	1962	ongoing
Commonwealth Secondary Scholarship Examination (CSSE)	1964	1973
Tertiary Education Entrance Project (TEEP)	1967	1970
Australian Scholastic Aptitude Test (ASAT)	1970	1993
From Testing to Teaching in the Humanities by Grant and Withers published	1973	1973
Australian Law Schools Entrance Test (ALSET)	1978	ongoing
History of ACER by Connell published	1980	
AST writing test developed	1985	ongoing
Special Tertiary Admissions Test (STAT) – developed from special versions of ASAT	1982	ongoing
Graduate Medical Entry Test (GAMSAT)	1992	ongoing
General Achievement Test (GAT)	1993	ongoing
ACT Scaling Test (AST) – developed from ASAT	1994	ongoing
Queensland Core Skills Test (QCST) – developed from ASAT	1994	2018
Aptitude Profile Test Series (APTS) – current release	2000	
A Historical Perspective on the SAT 1926–2001 by Lawrence et al. published	2002	
ACT Short Response Test developed	2004	ongoing
National Assessment Program – Literacy and Numeracy (NAPLAN)	2008	ongoing

1

A preamble

This monograph looks back 50 years to the beginning of a distinctive kind of testing that was developed and has taken place since that time at the Australian Council *for* Educational Research (ACER). It is hoped that the reader will see in this discussion why the test battery that began it all is worth remembering.

The story begins with the decision of the Menzies Government to award 10,000 scholarships for Years 11 and 12 to students who performed best in an examination at Year 10. The **Commonwealth Secondary Scholarship Examination** (CSSE) evolved into the first Commonwealth funded national examination in Australian history. The first version of the CSSE was developed by ACER and administered on the 7th of October 1964. Over the next decade the CSSE grew until around 100,000 students took it each year.

The development of the CSSE shaped the subsequent development of the **Cooperative Scholarship Testing Program** (CSTP) and the **Australian Scholastic Aptitude Test** (ASAT) by ACER in the 1970s. These programs have been running now for more than 40 years, and they are the oldest projects still in operation at ACER. 2014 was the 50th anniversary of the first version of the CSSE test, and along with it, those teenagers who sat the first test in 1964 were reaching their mid-60s.

In his 1980 history of ACER, W.F. Connell quotes the Secretary of the Commonwealth Department of Education and Science as referring in 1967 to the 'ACER-type' of testing. This monograph is the story of the 'ACER-type' of generic, cognitive skills testing that was developed in the 1960s, and is still evolving in the second decade of the 21st century.

This ACER approach to testing grew out of some quite specific promptings, and with time it has evolved to take on other roles and purposes. The 'ACER style' of testing came to underpin the school-based assessment systems that developed, with assistance from ACER, in Queensland and the Australian Capital Territory (ACT) in the 1970s, and that continues into the present.

Overview

The 'ACER style' of testing is discussed in this monograph under the more definitive title of '**cross-curricular testing**'. The meaning and significance of this title is considered during this discussion.

There are many aspects to the story of the 'ACER style' of testing. There are the socio-political forces that shaped the CSSE program, and the educational significance of the subsequent use of

'ACER style' tests in Queensland and the ACT. There are the technical and psychometric issues involved with cross-curricular testing, the controversial aspects prompted by this style of testing, and there are the personal reminisces of those who have worked on these tests.

While all of these aspects of the history of cross-curricular testing at ACER are worthy of attention, this discussion focuses on the notion of a cross-curricular test as distinct from other kinds of generic skills tests. It helps explain why the 'ACER style' is best described as cross-curricular, and it puts that approach in the context of a developing concept of intellectual abilities within cognitive psychology over the past 120 years. As well as dealing with the view of cognitive abilities implicit in the ACER cross-curricular testing style, this discussion considers the way in which cross-curricular tests have evolved, in particular the evolution of the CSSE into the ASAT and the present **ACT Scaling Test** (AST).

After the initial conceptualisation of the CSSE (discussed in Chapter 3) and the activity around the **Tertiary Education Entrance Project** (TEEP) and the early forms of the ASAT, there has been a paucity of conceptual analysis of the cross-curricular tests developed by ACER over the last 50 years. This monograph aims to undertake a comprehensive review and conceptual analysis, with the view of preserving some of the history for future test writers.

Chapter 2 of this text presents an adapted version of Bill Connell's discussion of test development at ACER up to 1980, and continues that discussion of cross-curricular tests to the present. The aim of this discussion is to both extend and enrich Connell's discussion of the 'ACER-type' of testing by examining that style in detail, and by placing it in the history of cognitive testing in the 20th century.

Understanding the distinctive style of CSSE, CSTP, TEEP, ASAT and the tests that followed in that tradition depends on seeing those tests against the backdrop of the development of cognitive tests in general. So this story involves a comparison of the CSSE tradition with cognitive testing in general, and the SAT test (once but no more called the Scholastic Aptitude Test) of the College Entrance Examination Board in the United States in particular.

A thread that runs through this discussion is the way the ACER-type of test sought and elicited sophisticated test items, and also involved delicate questions of balance in constructing a whole test. It is these aspects of cross-curricular test development at ACER that prompt the reference to 'the art of item writing' (and the concomitant art of test construction) in the title of this text. Among other things, this is the story of how CSSE encouraged the writing of subtle and penetrating test items that raised item writing and test construction to an art form. Through this discussion of specific groups of items (or units in ACER parlance) it is hoped that the reader will see what it means for item writing to be an art.

There were (and still are) some not insubstantial differences in the curricula of different Australian states and territories which would have made a curriculum-based examination for the award of scholarships difficult in the 1960s. So rather than setting out to test what students had learned, the designers of the CSSE set out to test the generic thinking skills that underpin and are developed by schooling. Rather than testing knowledge of different curriculum areas, the test focused on the higher-order thinking skills needed for learning in schools, and that were thought to indicate the ability to learn in the future.

The CSSE testing was stimulus-based rather than knowledge-dependent. Rather than testing what a candidate knew when they came into the testing venue, CSSE set out to test what candidates could do with the stimulus material presented to them during the test.

There were three separate multiple-choice tests in the CSSE called the Comprehension and Interpretation (Humanities), Comprehension and Interpretation (Sciences), and Quantitative Thinking (that also came to be known as Comprehension and Interpretation (Mathematics)). These

tests largely consisted of multiple-choice items and they were each of two hours duration. There was also a two hour test of Written Expression in the CSSE. The stimulus used in the different multiple-choice question tests was oriented towards different curriculum areas, and the stimulus was offered to candidates as meaningful challenges to the thinking process.

The tests were developed using psychometric techniques in that items were trial tested in New Zealand and placed in tests with the help of statistical analysis of the performance of the trial candidates. Although the test was developed on the basis of trialling and psychometric analysis, it was not based on and did not test conventional psychometric constructs. How the CSSE contrasts with conventional psychometric constructs is described in Chapter 4.

The CSSE was an extraordinarily broad, rich and comprehensive test battery. It differs with other tests developed elsewhere in the 1960s and used for similar purposes. The progeny of CSSE and ASAT still contrasts markedly with other tests used for similar purposes in other countries. This difference is explored through a comparison of the SAT with ASAT in Chapter 5.

While the CSSE had four separate tests, the results of them were to be aggregated. Over time the tests became less differentiated and more generic. In the ASAT the sub-tests were part of an integrated whole that could be separated into component sub-tests. There was an attempt to balance considerations of breadth and homogeneity in the design and construction of the ASAT. Over time, the skills tested became less domain-specific and the overall integration of the ASAT battery was clearer.

The ASAT battery evolved so that its progeny, the current AST, is very broad, but it has a reasonable degree of internal consistency as a whole. The distinctive characteristic of the ACER cross-curricular tests is that they balance meaningful generality with a degree of domain-related differentiation.

As a whole, the CSSE and ASAT are tests of general cognitive ability, but they can be partitioned into two (or more) very broad domain-related segments. These components are not envisaged as separate abilities, they are the core reasoning abilities conditioned by different topics and modes of representing ideas and information. ACER cross-curricular tests envisage general and very broad abilities as taking on domain-related characteristics and colouration.

The ACER cross-curricular tests contrast with intelligence, differential ability and aptitude tests that are based on psychometric constructs. ACER cross-curricular tests are focussed on integrated and higher-order thinking skills rather than a conception of separate and basic thinking skills. This difference is explored in Chapter 6.

Chapter 7 traces the evolution of writing assessment at ACER from the CSSE test of Written Expression into the current writing tests of CSTP, ASAT, STAT and GAMSAT, and Chapter 8 considers the evolution of the CSSE test battery, based on multiple choice items and extended writing, into a range of constructed response items in the current ACT Scaling Test.

Chapters 9 and 10 look at the art of cross-curricular item writing through examples from CSSE humanities, mathematics and science sub-tests. Various test items offered as examples in the text are analysed in detail.

Chapter 11 provides a detailed examination of the evolution of one piece of stimulus material in various multiple-choice items, which are in turn compared with constructed response items about the same stimulus material. This discussion aims to consider in specific terms the pursuit of significant and sophisticated test items from the point of view of the item writer.

Types of test construct

There are a number of ways in which examinations and tests can be understood. Assessment instruments can aim to test knowledge, or they can aim at specific or general skills. They can involve understanding acultural and abstract or on domain-related and contextualised ideas and information.

A knowledge-based assessment is a subject examination or an achievement test. Some other assessments focus on general abilities rather than knowledge or domain-specific skills. A generic skills test assesses an integrated higher-order complex of abilities that are designed primarily for overall assessment, but can be partitioned into useful components to some degree.

ACER cross-curricular tests have been defined in terms of the:

- subject or *topic* to be understood and analysed;
- *mode* of representing ideas and information in the stimulus; and
- *kind of thinking* involved in understanding and analysing the ideas and information.

The *topic* and *mode* are quite definite elements of a cross-curricular test, but *kind of thinking* is an elusive notion that is more difficult to identify and define. Breadth and balance in a cross-curricular test is a matter of informed judgement rather than explicit specification. A test specification cannot define the breadth and balance of a cross-curricular test, and it is possible to produce items in a nominal category that are quite different in tenor and texture. Cross-curricular tests can have the required number of mathematics, science or humanities items and be quite different in actual breadth and balance.

The ACER-type of cross-curricular tests are skills rather than knowledge based. The skills tested are generic rather than specific, and have domain-related, real-world characteristics rather than being abstract or acultural.

Purposes for cross-curricular tests

The ACER-type cross-curricular tests are very versatile, and they have been used for a range of purposes. The CSSE and CSTP have been used to identify academic aptitude and predict ability to learn in the future using the total score for the award of scholarships. The ASAT and the **General Achievement test** (GAT) have been used to correlate with individual school subjects or groups of subject assessments, The AST cross-curricular tests have been used to put different subject scores on a common scale, and to adjust the scores of different schools in a system for the construction of a single tertiary entrance score. ACER-type cross-curricular tests have been used as quality control for examinations and school assessments, and they have been used to monitor school and teacher effectiveness in value adding analyses.

This discussion does not deal with the psychometric characteristic of cross-curricular tests or the various uses made of such tests. It is concerned instead with the nature and characteristics of the test items themselves. It compares them with other cognitive tests to show what is distinctive about them, and to claim that such tests are authentic assessments of higher-order thinking.

Evidence for these claims are offered by examples of stimulus material and questions from various cross-curricular tests. The reader may glance at these examples and move on, but it is hoped that the examples and the comments that accompany them are worthy of some detailed attention.

Cross-curricular test items and whole units can differ spectacularly in quality. There is no need to offer examples of superficial, tedious and specious test items, but there is reason for offering examples of subtle, intriguing and penetrating test items, and drawing attention to their strengths. It is the inevitable fate of the item writer as artist to feel that the quality of their best work goes unrecognised. Or to find that their best ideas ‘don’t work’ and are binned.

The following page is an example of the unusual intellectual challenges to be found in some CSSE material. It is based on a surprising and imaginative process of matching drawings and poems (see Exhibit 1). The grounds on which such a matching can be made are not declared. Candidates are expected to feel their way with these comparisons, as it were. Item 56, for instance, asks candidates to choose a piece of poetry that best matches the *Standing nude* drawn by Pablo Picasso in cubist style in 1910.

Examples of CSSE testing

Exhibit 1 from a CSSE Humanities unit of 1971 asked candidates to compare five painting with five poems. They were directed to note 'the different types of lines used in each, the differences in the arrangement and effect created in the six drawings'. They were also directed to make 'a similar comparison of the poetry extracts' and note 'the total effect created, the way words have been used in the lines, and the rhythms created by them'. The candidates were given an example.

> *'For instance, Drawing V can be matched with The Tide Rises, the Tide Falls by H. W. Longfellow. The rise and fall of the rhythm and the smooth fluency of the poem match the smoothness of the lines and the texture they create in the drawing.'*

Specifically, Item 56 asked candidates to match a cubist nude by Picasso with four poems or extracts. This item (and the unit as a whole) is a very adventurous example of multiple-choice.

In 1973 a text was published by ACER called *From Testing to Teaching in the Humanities* in which Audrey Grant and Graeme Withers discussed the CSSE Humanities units of 1971 on which they had worked. Grant and Withers (1973) explained Item 56 in the following terms.

> *The instructions given on the paper directed pupils to look at the arrangements and effect of the lines used in the drawings, and at the words and rhythms of the poetry extracts. In addition candidates were asked to consider the total effect of each piece of work, poem, or drawing.*
>
> *Question 56, based on Drawing II, 'Standing Woman' by Picasso, had as its answer the extract from Eliot's 'The Waste Land'. Forty-five per cent of candidates chose correctly. Of the four possibilities, this extract is the one which matches best in tone and construction the effects of austerity and anonymity in the Picasso drawing. In both pieces smoothness of line or rhythm, to be found in the other three extracts offered as alternatives, is absent. Much of the individuality of Eliot's landscape and Picasso's figure has been ignored; and in both, only certain physical attributes remain, expressed jaggedly or irregularly in broken angular groups of lines. When one looks at A, B, and C, one can note a regularity or at least fixed quality of rhythm and a fluency of expression which do not match Picasso's composition as well as do Eliot's abrupt statements. (p.41)*

Exhibit 1.
Matching drawings and poems, CSSE 1971
Unit 5, Item 56

56. Which one of the following best matches Drawing II?

A. Dream within a Dream
I stand amid the roar
Of a surf-tormented shore,
And I hold within my hand
Grains of the golden sand —
How few! Yet how they creep
Through my fingers to the deep
While I weep — while I weep!

Edgar Allan Poe

B. The Tuft of Kelp
All dripping in tangles green,
Cast up by a lonely sea,
If purer for that, O Weed,
Bitterer, too, are ye?

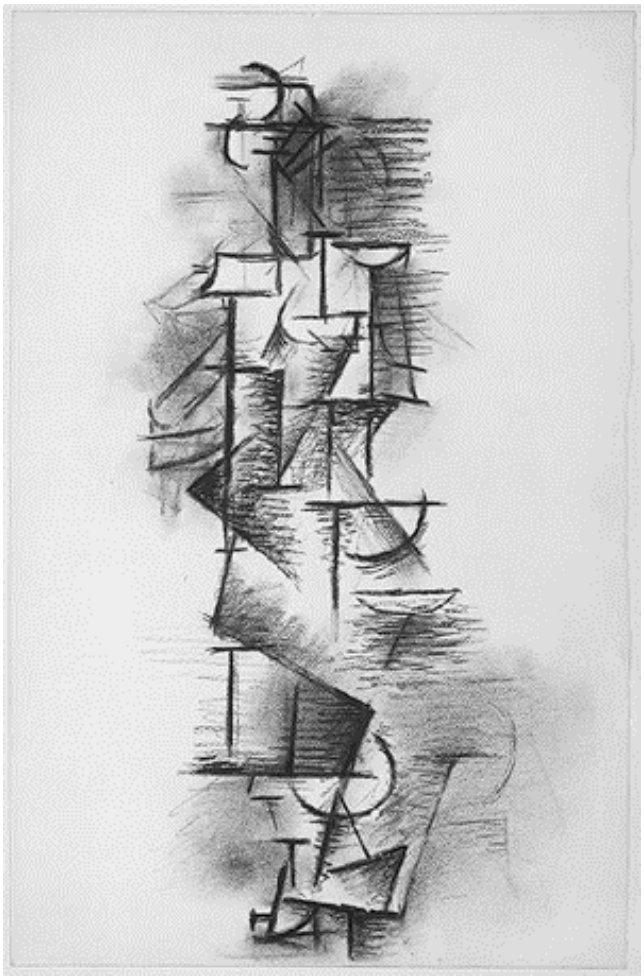
Herman Melville

C. The Sea
I have ridden the surf with foam across my face
I have rolled like a warrior of old
down the sea-wind in my chariot of water.
I have lain in the sea at twilight
as in a bed heaped up with flowers
clothed in my robe of water like a king.

A. R. D. Fairburn

D. And no rock
If there were rock
And also water
And water
A spring
A pool among the rock
If there were the sound of water only
Not the cicada
And dry grass singing

T. S. Eliot



Drawing II Picasso Standing nude 1910

The CSSE Science test went to considerable trouble to find interesting stimulus material, such as the cartoon strip on aspirin, presented in Exhibit 2, and Exhibit 3 from the Center for Short-lived Phenomena of The Smithsonian Institute.

The Aspirin Age unit (see Exhibit 2) uses a strip cartoon to tell the contemporary story of research into aspirin by John Vane which won him the Noble Prize in 1982. The material is concerned with the exploratory nature of science, and deals with unresolved questions about how aspirin works. The comic book presentation gives a drama to the story. There is an understated heroism about the presentation of Vane. You see Vane at his microscope and with patients and colleagues. The science is forward looking. What will aspirin and prostaglandin mean for contraception and the challenge of population growth?

Exhibit 2.
The Aspirin Age, Science 1972
Items 1-6

The drawings and information presented under the title 'The Aspirin Age' are to be used in answering Questions 1-6.

1. Which of the following is not given as an effect of aspirin?
A. Inhibition of prostaglandin production.
B. Reduction of inflammation.
C. Lowering of body temperature.
D. Curing of common colds.*

2. Aspirin was first observed to be a pain reliever
A. by the Rev. Edward Stone in 1763.
B. by German doctors in 1900.*
C. by Professor Vane during his work on histamine.
D. when it was related to male and female reproductive functions.

3. Research has not yet shown that prostaglandins are involved in
A. inflammation. C. fever.
B. reproduction.* D. pain.

4. Of the following, which have been shown to be involved in inflammation reactions of the body?
A. Histamine only.
B. Prostaglandins only.
C. Both histamine and prostaglandins.*
D. Acetylsalicylic acid only.

5. To date, research in this area has revealed that
A. what prostaglandins cause, aspirin can cure.
B. aspirin plays a key role in birth control methods.
C. prostaglandins can be synthesized from acetylsalicylic acid.
D. aspirin may relieve inflammation by preventing certain substances reaching the site of injury.*

6. Which of the following best describes the present status of aspirin in the eyes of the scientists?
A. Pharmacologists have tried in vain to discover how it works.
B. An understanding of the way aspirin works has resulted in the production of new aspirin-like drugs.
C. Current research has led to a partial understanding of how aspirin works.*
D. In taking 4,000 million tablets per year, Britons have provided an excellent experimental population for verifying aspirin's safe and sure effects.

Exhibit 3 is about the famous Murchison meteorite shower of October 1969 named after the town of Murchison in Victoria. It is one of the most studied meteorites showers ever because of its large mass (>100 kg), the fact that it was an observed falling, and that it belongs to a group of meteorites rich in organic compounds.

The stimulus is the reports about the meteor shower by Mr R.O. Chalmers of the Australian Museum, Sydney (on the 9th of October 1969) and Dr J.F. Lovering of the Australian National

University (on the 10th of October 1969). With remarkable agility, these first hand reports were turned into CSSE testing material by July 1970. The material was first hand, local and significant science in real time.

The questions ask about the intention and the substance of Lovering's report. Item 37 asks about differences between the two reports, and Item 38 asks about the way the accuracy of Lovering's report might have been compromised. The questions are first hand critical thinking about scientific observation.

Exhibit 3.
The Center for Short-lived Phenomena, Science 1970
Items 36-40

Read quickly through the following material (on pages 15 and 16) to gain a general impression; then look at the questions before turning back to the material.

The Smithsonian Institution in the USA maintains a 'Center' for recording and reporting on short-lived natural phenomena. Below is some information about the Center.

SMITHSONIAN INSTITUTION
CENTER FOR SHORT-LIVED PHENOMENA

INFORMATION ABOUT THE CENTER

The Purpose of the Center

The purpose of the Center is to act as an early-alert system and clearing house for the collection and dissemination of information on short-lived natural phenomena. The Center uses an international communications system capable of reaching virtually any point on earth within minutes. This communications ability enables the Center to contact event areas quickly to obtain information on events while they are occurring or shortly after they occur.

Correspondents of the Center

The Center has established a global network of more than 2000 scientific correspondents in 122 countries. Correspondents are scientists, scientific institutions and field stations that cooperate with the Center by reporting short-lived events that occur in their areas and in return receive reports of interest to them from the Center.

Event Notification and Information Cards

The Center notifies subscribers of the occurrence of short-lived events through the issuance of Event Notification and Event Information cards. These cards are issued within 24 hours of the receipt of the event information by the Center and are immediately dispatched to subscribers. Event Notification cards contain initial information on the occurrence of current or recent events. Event Information cards contain additional information and data and the current status of continuing events and/or preliminary results of field expeditions.

Two reports of a meteorite shower in Victoria were received by the Center and sent to interested persons. The reports are reproduced exactly as they were received.

Event	**125-6**	**Victoria Meteorite**	**9 October 1969**	**783**
The following table report was received by CFSLP on 9 October 1969: METERORITE BREAKING UP OVERHEAD NEAR MURCHISON, VICTORIA. LATITUDE: 36' 37' SOUTH LONGITUDE: 145' 12' EAST TIME: 1 15 Am 28th SEPT. PUFFS OF SMOKE AND EXPLOSIONS. TRAVELLING FROM SOUTH EAST TO NORTH WEST. COMPLETE STONES AND FRAGMENTS FELL OVER 5 MILES ALONG A MILE WIDE TRACK EXTENDING FROM 5 MILES SOUTH EAST OF MURCHISON. MANY PICKED UP BY LOCAL RESIDENTS. THE STONES ARE CARBONCEAOUS CHONDRITE. I'VE GOT 2 SMALL STONES. PROFESSOR J.P. LOVERING, GEOLOGY DEPT, UNIVERSITY OF MELBOURNE IN CHARGE OF INVESTIGATION. HIS TEAM COLLECTED OR BORROWED A NUMBER OF SPECIMENS INCLUDING THE LARGEST SO FAR KNOWN. AN INCOMPLETE MASS OF 1.5 LB.			Event information	
			Type of event	Astrophysical
			Date of occurrence	28 September
			Location of event	Murchison
			Victoria, Australia	
			Reporting source	
			The Australian Museum, Sydney	
			Source contact	Mr R. O. Chalmers
			The Australian Museum Box A 285	
			SMITHSONIAN INSTITUTION CENTER FOR SHORT-LIVED PHENOMENA Cambridge Massachusetts	

Event	**125-6**	**Victoria Meteorite**	**10 October 1969**	**785**
1) The following report by Dr Lovering was telephoned to CESLP. 'The meteorite definitely is 4 chondritic meteorite of type 2 or 3 carbonaceous chondrite. It's fairly compact in most places but it's a friable in certain areas and appears to be at least somewhat heterogeneous. The correct time for the fall was sometime during the period of 10:45 a.m. to 11:03 a.m. The area of the fall is five miles by one mile. The stone fragmented in the air. The fractured surfaces have incipient fusion crusts on them. A strong smell was reported associated with the meteorites when they were recovered. I didn't believe this but checked it out and found that indeed they had the odor of methylated spirits (denatured alcohol). I have collected and have in my possession some of it on loan, 2.5 kg. of material. I know of another 4.5 lbs., approximately, in the hands of private individuals. It's a total of about ten pounds that has been recovered so far.'			Event information	
			Type of event	Astrophysical
			Date of occurrence	28 September '69
			Location of event	Victoria
			Australian National University	
			Reporting source	
			Australian National University	
			Source contact	Dr J. F. Lovering
			Dept. of Geophysics and Geochemistry ANU PO. Box 4 Canberra ACT 2600	
			SMITHSONIAN INSTITUTION CENTER FOR SHORT-LIVED PHENOMENA Cambridge Massachusetts	

36. The first part of Dr Lovering's report was intended to
A. classify the meteorite according to a scientifically accepted system.
B. name the meteorite so that scientists could distinguish it front other meteorites of the same kind.
C. give a theory of formation of the meteorite.
D. suggest that the meteorite was formed from coal.

37. Of the following, the major point of disagreement in the two reports was concerned with the
A. composition of the meteorite.
B. reported time of sighting.
C. amount of material collected.
D. area and position of fall.

38. If Dr Lovering had delayed his arrival for another week, which of the following may not have been accepted by him as a fact?
A. The material was carbonaceous chondrite.
B. Fractured surfaces bad incipient fusion crusts.
C. The material had an odour of denatured alcohol.
D. The material was heterogeneous.

39. Which of the following descriptions is consistent with that of the meteorite material described in the reports?
A. A hard stone with particles of the same size having patches of colour.
B. A hard stone with patches of crumbly material containing particles of the same size.
C. A black, coal-like stone with a uniform glassy appearance.
D. A stone with patches of crumbly material containing particles of different size.

40. Which of the following events is most likely to claim the attention of the Smithsonian Institution?
A. A report that an earthquake struck a village in the centre of the Philippines.
B. A report that the buildings of Venice are steadily sinking into the sea.
C. A report that a small tribe of natives in South America use a poison which causes death within a few minutes of its coming in contact with broken skin.
D. A report that a chemist has produced a sample of a new element, the sample existing for less than a millionth of a second.

2

The history of cross-curricular testing at ACER 1960 to 2014

From its beginning in 1930, ACER has produced the standard general and differential ability tests of the psychometric tradition. ACER produced the first non-verbal test of general ability for Australia in 1936. The most recent test of this kind is the **Aptitude Profile Test Series** (APTS) released in 2000. A list of some of the general ability testing material produced by ACER since the 1930s and those that are currently sold can be seen in Appendix 1. Some of the tests developed in the 1930s and 1940s are still used in some form or another. Accordingly, this chapter is based on and extends Chapter 9, Testing and Research Programs 1960 to 1980, of the history of ACER, by W. F. Connell entitled *The Australian Council for Educational Research 1930-80.*

The tests mentioned in Appendix 1 are more or less conventional intelligence or differential ability tests adapted from overseas models. In contrast with this kind of testing material, another major and distinctive kind of testing material was developed by ACER in the 1960s. These tests are best described as 'cross-curricular' because they test general skills, but the skills tested are those that underpin and reflect broad educational domains. This style of cross-curricular testing developed in programs for the award of scholarships and selection into tertiary institutions.

ACER first produced general ability tests for the award of educational scholarships in the 1940s. It was also involved in the following projects and programs for selection into tertiary institutions in the 1940s and 1950s.

- Examination of University of Melbourne students for admission under adult matriculation scheme (1940-55)
- Testing of applicants for entry to Melbourne Technical College (1942-45)
- Testing of teachers college entrants (1955-59)

In the 1960s a new era in testing began for ACER with programs for various states and territories, and subsequently for the Commonwealth Government. The key tests in the evolution of cross-curricular testing at ACER were the **Commonwealth Secondary Scholarship Examination** (CSSE), the **Tertiary Education Entrance Project** (TEEP), and the **Australian Scholastic Aptitude Test** (ASAT).

ACER tests for the award of scholarships and selection into courses

In the 1940s ACER was involved with the examination for Victorian Junior Scholarships awarded for secondary school courses. From 1945 to 1960 ACER developed or modified existing tests of general ability for students of 12 to 13 years of age who sat for that scholarship examination. In the mid-1950s several experiments were started with the use of tests of scholastic aptitude and skills in reasoning in the selection of students for entry to various tertiary institutions. This

scholarship and selective testing developed into the distinctive style of ACER cross-curricular testing.

The Cooperative Scholarship Testing Program

In 1962, a number of independent schools in Victoria who offered scholarships at a similar age level requested ACER to supply them with a test package including tests of English, mathematics, and general ability for use as a common examination. In this program ACER marked the papers and provided all schools with a list of students in order of merit. In the following year, several independent schools in New South Wales joined the group and the program soon spread to all states and territories. In 1972 the tests were substantially revised in an effort to make them more a measure of scholastic aptitude rather than achievement. They were offered at three levels, one for scholarships at entrance to secondary school, and two others for continuation beyond Years 8 and 10. The battery contained tests in the humanities, written expression, and mathematics-science, as a somewhat junior version of the CSSE (for which the ACER had also been responsible for the previous eight years). The **Cooperative Scholarship Testing Program** (CSTP)[1] has continued to the present time. The CSTP has regularly involved more than 100 independent secondary schools in all states and territories of Australia and approximately 12,000 students each year in the 1970s.

The Commonwealth Secondary Scholarship Examination

In 1963 the Commonwealth Government decided to award 10,000 scholarships for students going into Years 11 and 12 – the last two years of secondary school. The scholarships were to be awarded on the basis of a competitive examination.

In Victoria there was no external examination at the end of Year 10 that could be used for the purpose of awarding the scholarships. In 1964 the Victorian Education Department requested ACER to develop an examination in several subject areas, not associated with the content of a set syllabus, on which the new set of scholarships might be awarded. ACER undertook the project and proposed that there be test papers in four areas: Written Expression, Quantitative Thinking, Comprehension and Interpretation (Sciences), and Comprehension and Interpretation (Humanities).

ACER recommended that the tests focus on the central areas of secondary education. By examining the intellectual skills in all four areas, the project would be testing whether students had a grasp of the basic skills needed to complete secondary education. The Director of ACER, W.C. Radford, described the tests in the following way.

> *The abilities which they would test are those which should be developed in any good broadly based course of study, irrespective of its specific content. . . . They would test, as a whole, the range of abilities which we think a soundly educated child should have at this level, i.e. [the child] should have a good background in literature, social studies, mathematics, and science (a background which has produced abilities which are not restricted to knowing or applying a particular content of a particular course, and which are transferable to new and different courses), should be able to write well, and should know something of the mechanics of expression. These are abilities which should be developed in any kind of school.*

Radford's comments expressed a new aim for general ability testing and a new and different relationship between ability testing and the school curriculum.[2] These aims came to shape the

[1] In the 1970s the CSTP battery was revised to involve two writing tasks and tests of comprehension and interpretation in mathematics and science and the humanities and social sciences, and so it remains to the present.

[2] From *The Objectives Being Tested in the Commonwealth Secondary Scholarship Examination* prepared in 1967 by T. M. Whitford presented in Appendix 2.

distinctive approach to cross-curricular testing taken by ACER in the 1970s and continued into such current tests as **the Cooperative Scholarship Testing Program** (CSTP), the **ACT Scaling Test** (AST), the **Special Tertiary Admissions Test** (STAT), the **Australian Law Schools Entrance Test** (ALSET), the **Graduate Australian Medical School Admissions Test** (GAMSAT) and the **General Achievement Test** (GAT).

Once the work on the CSSE test battery got under way, the other Australian states and territories joined the program.

There was some previous experience to help the test constructors in the work that had been done by ACER for the entrance scholarships for independent schools, but the new work was pitched at a higher level and covered a wider range. It was a 10 year project. At the end of 1964, ACER signed a three year contract with the Commonwealth Government, which was renewed until 1974 when the scholarship scheme was discontinued after the election of the Whitlam government.

The information presented in the tests was drawn from literature, historical documents, social studies tables, cartoons, geographical maps, photographs and diagrams from scientific and psychological experiments, as well as from mathematical patterns that include numbers, algebraic symbols and line drawings. An extensive use was made of objective test items and great care was taken in the construction of them. The tests were subject to close scrutiny by teachers and other educational research workers. The test construction team consequently spent much time on refining the material and the questions so that they would withstand criticism from such sources. The head of test development at ACER wrote of the test that: "We feel a special responsibility in this area as the producers of a major objective test battery which functions as a model for Australian education" (Clark & Ennor, 1970, cited in Connell, 1980).

The test development work of ACER in the CSSE was stimulating to the teaching profession (Grant & Withers, 1973). It is probable that the teaching and examining in English during the 1970s was affected by the test papers in written expression and the humanities, which were widely studied by Australian educators.

Some of the material in the test was a little controversial. The test constructors were accused of blasphemy on one occasion when an unusual translation of the Bible was used in a question, of immorality for using reproductions of nude paintings, and of subversive communism when another question featured Mao Tse-tung.

Tertiary Education Entrance Project and Australian Scholastic Aptitude Test

In 1967, the Commonwealth Department of Education and Science raised the possibility of evolving an 'ACER-type' test (that is a CSSE-type test) as a university entrance examination. The Commonwealth Department commissioned ACER to undertake a two-year investigation using CSSE-type tests. The **Tertiary Education Entry Project** (TEEP) was launched towards the end of 1967, and had the general aim of reviewing various methods of assessment for tertiary education with a view to improving selection.

ACER was required to prepare five papers on the CSSE model for use in 1968: one essay-type paper of written expression, and four papers of objective multiple-choice questions on quantitative thinking, comprehension and reasoning in the physical and biological sciences, comprehension and reasoning in the social sciences, and understanding and interpretation in the arts and humanities. It was the expectation of the test constructors that the tests would be a good predictor of success in tertiary education and would have sound educational value in directing teachers' attention to the kinds of questions they should be asking their students.

In 1969, ACER was commissioned, as the second stage of the project, to prepare a different kind of test, a 3 hour multiple choice test of scholastic aptitude, for experimental use in 1970. The **Australian Scholastic Aptitude Test** (ASAT) was administered in all states and territories except Victoria at the end of 1970. The TEEP tests were also used in that year in various states and territories. The Commonwealth Government continued to fund the production of ASAT until 1974, at which stage two states and two territories (Queensland, Western Australia, Northern Territory and the Australian Capital Territory) were using the test. Further versions of the test have been paid for by the users.

The ASAT was an attempt to assess a range of intellectual skills that were thought to be important for students in tertiary-level work. Like the CSSE and TEEP tests, it was composed around various kinds of stimulus material not taken from prescribed school syllabuses. It was intended that, when combined with an assessment of a student's school work made by teachers at the school, the combination would predict success in tertiary studies at least as well as the traditional external examination. In its first version, it was a three hour test; later a four hour test became available. The production of ASAT has continued in various forms to the present time. After the first few experimental years, ASAT was regularly used only in Queensland, Western Australia, the Australian Capital Territory, and the Northern Territory. From 1973 until the early 1980s, special forms of ASAT were produced for institutions that wished to select among candidates who did not fit into the normal entry categories. This work developed into the **Special Tertiary Admissions Test** (STAT).

As the users and constructors of ASAT became more familiar with it, both its educational possibilities and its use in scaling other assessments became more apparent. From 1973 on, several research studies were undertaken (similar to those previously done in connection with CSSE) from which it was possible to make continuous improvements in the structure of the various forms of the test.

After 15 versions of the test, the Board of Senior Secondary Studies in Queensland decided in 1990 to develop its own version of the ASAT to be called the **Queensland Core Skills Test** (QCST). ACER test constructors were consultants in test design and item development to the Board between 1990 and 1994. The QCST is currently in use as the only external assessment in the Queensland system of school-based assessment. Test scores are used to place the results in different subjects and the results in different schools on the same scale to make them comparable.

This system of using a test battery to statistically moderate school-based assessment was proposed by the Director of ACER in the 1970s, and it has been used as part of school-based assessment systems in Queensland and the Australian Capital Territory ever since.

The ACT Scaling Test

The ACT Scaling Test (AST) is the current version of ASAT that has been especially developed for the Canberra education system of school-based assessment. The battery includes a 135 minute multiple choice test of the general skills that underpin the mathematics, science and technology domain and the humanities, arts and social sciences domain. The battery also includes a two and a half hour writing test (discussed in Chapter 7) and a 105 minute short answer question test (discussed in Chapter 8). The AST program in Canberra is used to moderate between subjects' scores within colleges and aggregate scores across colleges. Appendix 4 provides a description of the AST system, followed by the 2011 AST specifications in Appendix 5.

The General Achievement Test

In 1993 the Victorian Board of Studies requested ACER to trial an ASAT-like cross-curricular test for the purpose of reviewing the relationship of such an instrument to the school assessed common assessment tasks that were introduced into the **Victorian Certificate of Education** (VCE) in 1991. These assessment tasks required review, and in some cases remarking by the Board, making the process costly. It was suggested by a review of the VCE that a cross-curricular test might be used as a sifting mechanism to determine which school assessed tasks should be remarked by the Board. After a positive relationship between **General Achievement Test** (GAT) scores and remarking by external assessors was shown in a trial in 1993, all candidates undertaking the VCE were required to sit the GAT in 1994. Subsequent experience showed that the GAT had a strong relationship to the performance of students on the common assessment tasks, and that the GAT very effectively targeted school assessments that were too low or too high according to the remarking process. Concern about the integrity of the common assessment tasks led to the abolition of these school assessed tasks in 1996, but by that time the Board had found other uses for the GAT.

The Board began in the mid-1990s to use the GAT as a common measure with which to examine the performance of students in particular schools. The results for each school were placed on a common scale by the performance of the students on the GAT. This scaling was used to construct a 'value added measure' that compared the extent to which schools assisted their students to perform in the examinations. The comparison of schools on the basis of the 'value added measure' became a matter of public controversy. A review of the procedure recommended that the result of the value adding assessment not be published, and that other information about student performance should be published. The use of the GAT for other purposes was explored by the Board between 1994 and 2000. On the basis of this activity it was recommended that the GAT be continued. The Victorian Curriculum and Assessment Authority currently uses the GAT scores:

- to contribute to statistical moderation of school-assessed coursework;
- to review school assessments in school-assessed tasks;
- for special provision and derived examination scores; and
- for checking the accuracy of examination marking.

The GAT was designed to assess the cross-curricular skills that underpin VCE studies and the Victorian Curriculum and Standards Framework outcomes. It has retained the same format since 1995. The test is made up of two writing tests of 30 minutes each, and a two hour test of 35 **Humanities, Arts and Social Science** (HASS) and 35 **Mathematics, Science and Technology** (MST) multiple-choice items. The complete test is undertaken in one three hour session. Excerpts from the Student Information Bulletin for the GAT are found in Appendix 6.

As has been shown above, the AST and the GAT are direct descendants of the CSSE, TEEP and ASAT developed by ACER in the 1960s and 1970s. Various purposes and uses have shaped the distinctive style of the cross-curricular testing that ACER has undertaken since the 1960s.

The beginning, middle and the end of the first CSSE Humanities test of 1964

The following examples are the first, eighth, and the last units in the first version of CSSE Humanities test. The CSSE tests evolved significantly over 10 years, but these examples show that the test items were ambitious from the start.

Items 4, 5 and 6 in Exhibit 4, test reading comprehension, the staple of a text based test, but they do so in an unusually diverse and flexible way. Item 5 is concerned with the connotation of a particular word, but Items 1 and 2 are inferential rather than literal, and Item 4 is focused on

stylistic issues. Item 3 is evaluative, and is an unusual way of asking about the narrative arc of the passage. The first unit in CSSE Humanities in1964 was not the literal comprehension of most reading tests of that time.

Exhibit 4.
The Drone of the Tambura, Humanities 1964
Items 1-6

Take your seat in the unadorned hall, resonant with the unchanging drone of the tambura, the basic single stringed instrument. The dancer bows to you, stands motionless for a moment, then clicks her fingers lightly to the tabla player's tones which grow more and more irresistible. She comes forward and bends as though to kiss the feet of the guru, who is seated at your side. She waits silently for an instant, about to begin her dance. She begins ... When, at a sign from the guru, she stops, you do not know whether she has been dancing for half an hour or an hour. She could have gone on all day, with more and more intricate steps, seemingly improvised, never weary, never frenzied, never mechanical but on the contrary with increasing physical control and ever-growing rapture.

Georges Fradier

1. The dancing conveyed to the writer a sensation of

A. renzy.
B. timelessness.*
C. tedium.
D. turbulence.

2. The writer implied that the dancer

A. danced with mechanical precision.
B. was made to follow every whim of the guru.
C. acted in a frenzied and uncontrolled manner.
D. carried out her sequences as they had been rehearsed.*

3. This passage is notable chiefly for its

A. suspense.
B. sense of contrast.*
C. continuous action.
D. technical vocabulary.

4. The style of the passage can best be described as

A. graphic.*
B. mechanical.
C. romantic.
D. metaphorical.

5. When the writer said the hall was 'resonant', he meant that it

A. as filled with the sound of the tambura.
B. re-echoed the tambura's drone too loudly.
C. reproduced musical sounds poorly.
D. reverberated with the sounds of the tambura.*

6. If you were to draw a line to depict the pace of the passage, it would look approximately like which one of the following?

A.* B. C. D.

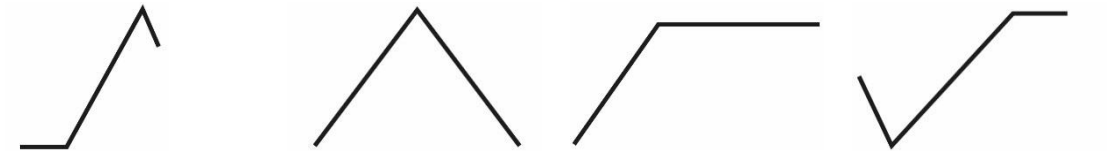


Exhibit 5 is a complicated piece of text by William Hazlitt. There are no literal questions about this 103 word passage. Items 43, 44 and 46 are inferences about the view of the writer. Item 45 is about the style of analogical thinking in the passage. Items 47 and 48 are about the rhetorical strategy and style of the language. These six questions are a model testing of inferential and meta-cognitive or reflexive reading.

Exhibit 5.
The Indian Jugglers, Humanities 1964
Items 43-48

I have always had this feeling of the inefficacy and slow progress of intellectual compared to mechanical excellence, and it has always made me somewhat dissatisfied. It is a great many years since I saw Richer, the famous rope-dancer, perform at Sadler's Wells. He was matchless in his art, and added to his extraordinary skill exquisite ease, and unaffected, natural grace. I was at that time employed in copying a half-length picture of Sir Joshua Reynolds'; and it put me out of conceit with it. How ill this part was made out in the drawing! How heavy, how slovenly this other was painted. W. Hazlitt, *The Indian Jugglers*

43. The writer shows concern over
A. his inability to perform on the tight-rope.
B. the comments of Sir Joshua Reynolds on his painting.
C. the lack of great thinkers at that time.
D. the comparatively slow improvement of the intellect.*

44. When the writer saw Richer's act, he
A. became disgusted with his own artistic efforts.*
B. considered Richer superior to Sir Joshua Reynolds.
C. was stimulated by Richer's powers of reasoning.
D. destroyed the work he had been doing.

45. The writer contrasts physical dexterity and
A. natural grace.
B. mental maturity.
C. mechanical excellence.
D. artistic proficiency.*

46. We may infer that, had Richer performed as the writer had painted, he (Richer) would probably have been
A. announced as a great discovery.
B. seriously or fatally injured.*
C. suspicious of his sudden success.
D. employed by Sir Joshua Reynolds.

47. The method employed by the writer is to
A. pretend to believe mechanical skill is of supreme importance.
B. present his case by analogy and reference to his own inadequacy.*
C. appeal to the emotions of his readers.
D. denounce manual skills as inferior to mental development.

48. The style of the passage may best be described as
A. ironic.
B. idiomatic.*
C. argumentative.
D. self-contradictory.

One suspects that the last unit in the first test, Exhibit 6, would have been significantly more difficult than Exhibits 4 or 5. Critical reading and interpretation is tested in the comparison of the view of Professors Oman and Davis. Candidates were given a general orientation in the introduction to the unit, but the language of both passages is complicated and particularly allusive. There are no footnotes, and candidates are expected to infer the significance of references about which one would expect they had no knowledge. They have to discern the general direction of an argument based on obscure references.

Exhibit 6.
Different views of the Norman Conquest, Humanities 1964
Items 88-93

The two passages on this page about the Norman Conquest of England in 1066 A.D. were written by Professors of History, both of them eminent and respected scholars. Read them carefully and then answer questions 88-100 which follow.

Professor I

Let anyone who believes that the Norman Conquest led in every sphere of civilization to a rapid and satisfactory development, compare the neat silver pennies of the later issues of Edward the Confessor and the short reign of Harold Godwineson, with the shapeless ill-struck issues of Henry I, Stephen, and the early years of Henry II. A relapse into barbarism might rather be deduced from the comparison. For good or ill the Conquest was accomplished - but the more we study it, the less easy it is to acquiesce in the easy and comforting conclusion that all was for the best - that the survival of an English England must necessarily have been a disaster. We are told that the insular Church and State were alike decadent, and the failings of Archbishop Stigand are held up for disapproval, along with the misdeeds of Earl Tostig. But Stigand is a less hateful figure than Odo of Bayeux or Ralf Flambard, on whose characters as typical Norman prelates anyone might dilate who wished to set forth the opposite theory.

The optimist may hold that the future development of this realm under continental influences was so infinitely superior to what that development would have been under purely national influences, as to compensate England in the end for all that she suffered in and after 1066. But the breaking up of the old governing class, the general confiscation of estates, the trampling of the nation beneath the feet of an alien aristocracy, were a heavy price to pay for that problematical gain. Episodes like William's ravaging of Northumbria in 1069, whose after effects endured for whole centuries and surpassed everything that the Dane ever wrought, cause us to doubt the theory that paints the Norman as the spreader of civilization. Were the tyranny of Rufus, the grinding oppression of Henry I, the anarchy of Stephen, necessary stages in the evolution of a nation? Can the introduction of Wager of Battle be considered a happy juristic reform? May it not be said that William the Bastard turned England from her true line of development towards the sea - she was a great naval power when he found her - and involved her in that unholy game of gambling for French provinces which was not to end till the Hundred Years' War was over, after four centuries of wasted effort?

(Oman, Sir Chas. England Before the Norman Conquest, pp. 649-51. *Methuen's History of England*.)

Professor 2

The Norman Conquest of England was the outcome of a struggle, short and spasmodic in character, between a handful of adventurers and a decadent nation lying on the outer fringe of European politics, and though it nearly affected the interests of several powers it occasioned no general disturbance of international relations. In fact, if the importance of an event were to be measured by the commotion which it makes among contemporaries, the Norman Conquest might be regarded as of little moment in European history. None the less it is one of these events which stands as a boundary mark between two stages of civilization and there is something more than accident in the rapidity with which, after the victory of Senlac, Europe emerges from the Dark Age into that splendid twilight which a large proportion of civilized humanity still prizes more highly than the morning light of the Renaissance or the mingled storm and sunshine of the Reformation. ... When Harold fell beneath the Dragon Standard, the last stronghold of Teutonic law and institutions, of a liberty which had degenerated into licence, of an aristocracy who had outlived their function and their virtues, was opened wide for the entry of the Italian priest and the Gallic legislator.

(Davis, H. W. C. England Under the Normans and Angevins, pp. 1-2. *Methuen's History of England*.)

88. In his second paragraph, Professor 1's purpose is to

A. show that the Norman Conquest brought in a period of barbarism.
B. cast doubt on the theory that ultimate benefits compensated England for her losses resulting from the Norman Conquest.*
C. combat the belief that after the Norman Conquest England quickly emerged from the Dark Ages.
D. suggest that in and after 1066 England suffered a great deal, but that in the long run English development benefited.

89. Professor 2 uses the term 'splendid twilight'

A. to emphasize his attitude that the Norman Conquest promoted Europe's cultural development.
B. to show that the Norman Conquest set the commencement of a period of backwardness.
C. to draw an analogy between England of the Norman Conquest and Europe during the Renaissance.
D. to imply that during the period following the Norman Conquest there was a general development throughout Europe.

For each of the statements 90 - 93 below, write in the answer space provided

A. if they are statements of fact which support the views of Professor 1.
B. if they are statements of fact which support the views of Professor 2.
C. if they are statements of fact which may be interpreted to either view.
D. if they are statements of an opinion more likely to be held by Professor 1 than by Professor 2.
E. if they are statements of an opinion more likely to be held by Professor 2 than by Professor 1.

90. William the Conqueror laid waste Northumbria. A

91. Before the Norman Conquest, England was in a state of barbarism. E

92. Archbishop Stigand's faults were less serious than those of Odo of Bayeux. D

93. Before 1066, England was a great naval power. D

The unit involved questions about the purposes of each passage, such as Item 88. Item 89 involved a difficult inference about the significance of a metaphor. Items 90 to 93 asked candidates to distinguish facts from opinions and attribute each to one or other writer, or to recognise that a fact might be adduced by both. This is difficult critical reading involving inferences and interpretation of a very high order. Such stimulus and questions were testing for Year 10 scholarship contenders. Such material would be very challenging for students in the 21st century.

3

What is a cross-curricular test?

The terminology of test constructs is sometimes used rather loosely, and it can be difficult to see what is being claimed (and what is being assumed) about what is being tested. Words can whirl around in the assessment business.

Words like 'knowledge' and 'skill' seem straight-forward, although when one begins to probe them they turn out to be very complicated. Things become even more difficult when we think about abilities, which are to be distinguished from competencies, which are in turn to be distinguished from capabilities. What is meant by terms like '*Verbal ability'* and '*Quantitative ability'* are also uncertain or variegated.

It is beyond the scope of this monograph (as well as the knowledge and skill of this writer) to determine the meaning of some key terms in assessment discourse, but an attempt must be made to clarify a few terms that are needed to tell the story of the evolution of cross-curricular testing at ACER. It is hoped that the process of definition is clarified by putting this conceptual discussion against the backdrop of the development of views of intellectual ability within cognitive psychology in Chapter 4. But first, a more conceptual delineation is appropriate, followed by a more concrete discussion of the distinctive kinds of cross-curricular tests developed by ACER.

A cross-curricular test as a whole is an assessment of the generic, central or core cognitive skills and abilities that underpin general education. Generic skills are the broad cognitive abilities or capacities used to learn novel tasks. These generic skills are distinguished from:

- the specific knowledge and skills needed for a particular performance;
- what the expert knows;
- competencies or an area in which you have to be competent; or
- a class of tasks.

Generic skills are best thought of as general abilities. They are the fluid capacities used by novices to gain expertise. The assessment of general abilities is focused on aptitude to learn, rather than achievement or competence. According to these definitions, generic skills are those abilities we use to learn new things as distinct from what has been learned in the past. A cross-curricular test of generic skills aims to be accessible to all reasonably able and informed candidates no matter what course of study they are undertaking.

While ACER's cross-curricular tests cover most or much of the common school curriculum, they are also thought of in terms of the two very broad curriculum areas of mathematics, science and technology (MST) and humanities, arts and social sciences (HASS). Cross-curricular tests are distinguished from subject-based examinations of curriculum, on one hand, and differential ability

tests of psychometric constructs like verbal, quantitative and abstract ability on the other. Cross-curricular tests are also distinguished from literacy (or literacies) in that they are defined through a focus on thinking skills and with reference to the topic and content being thought about. A cross-curricular test can also be distinguished from tests of more or less special abilities like problem solving, critical thinking and interpersonal understanding.

A cross-curricular test

A distinctive feature of cross-curricular tests is that they are stimulus-based groups of items rather than discrete item testing. A cross-curricular test puts some stimulus material (which is usually unfamiliar) before candidates and asks them to analyse and understand it. The stimulus will usually not come from text books or reflect specific subject area content, but it can have a relationship to different subject domains, and it will aim to test the kinds of thinking that typify broad subject areas.

The beginning of cross-curricular testing at ACER

When first introducing this new kind of testing in 1967, the second Director of ACER contrasted it with 'traditional general ability or intelligence tests'. (See Appendix 2 for the Objectives of the Commonwealth Secondary Scholarship Examination.) Dr Radford went on to write that

> *each paper attempts to test abilities which those at A.C.E.R. concerned with the examination believe will be developed by good teaching and interested learning, whatever the specific course of study. They are, we think, important abilities which are durable, and which should, if well developed, lead to success in later academic studies.* (Whitford et al., 1967)

The objectives for the Commonwealth Secondary Scholarship Examination offered the following rationale for this new kind of testing.

1. In the process of education (both formal and informal) a person develops skills and abilities as well as acquiring specific factual knowledge.
2. These developed abilities outlast much of the specific knowledge and are major objectives toward which a good teacher aims irrespective of subject course material taught in a particular school or State.
3. These developed abilities are important components of later academic success and it is therefore reasonable to use tests of such abilities to identify the most scholastically able students.
4. The most important of these developed abilities are considered to be:
   (a) Ability in the use of the mother-tongue, encompassing adequate vocabulary, recognition of correct and effective expression and the capacity to communicate correctly and fluently in writing.
   (b) Ability to read critically, to understand, to interpret, and to reason logically in the broad fields of the humanities and the sciences.
   (c) Ability to handle data presented in quantitative, symbolic or figural form.
5. It is possible to measure the development of these abilities through the medium of material which does not require the recall of specific information taught in subject courses.
6. Since tests based on the preceding propositions are concerned with the generalised outcomes of education and do not require teaching to a pre-determined 'subject-content' syllabus they do not restrict the freedom of the teacher. Further, such tests do not encourage

> memorisation of facts, formulae, procedures or mechanical application of standard formulae or procedures. Since they try to assess skills and abilities more important for later studies, any successful attempts at 'coaching' should result in relatively permanent educational gains as distinct from possibly temporary gains leading to success in the examination.

While the language used to describe and define the current cross-curricular tests developed by ACER has evolved in various ways, the spirit of the CSSE objectives developed in 1967 underpin such current cross-curricular testing programs as the **Cooperative Scholarship Testing Program** (CSTP), the **ACT Scaling Test** (AST), the **Special Tertiary Admission Test** (STAT), the **General Achievement Test** (GAT), and the **Graduate Australian Medical School Admissions Test** (GAMSAT Humanities and Social Science).

What does a cross-curricular test, test?

A cross-curricular test asks candidates to think about some stimulus material they probably have not thought about or been taught about before. A cross-curricular test aims to minimise the importance of knowledge and maximise the importance of reasoning skills in tested performance. It does so by aiming to test what a candidate can do with the stimulus material presented in the test, rather than testing what the candidate knows when they come into the testing venue. Cross-curricular tests are not based on psychometric constructs (as discussed in Chapter 4). Cross-curricular tests try to pose meaningful problems or challenge candidates to think analytically and critically about meaningful issues.

The ACER cross-curricular construct

The cross-curricular tests developed by ACER are oriented with reference to the curriculum of general education. A cross-curricular test aims to be a broad and balanced test of thinking skills. The notion that a cross-curricular test is a whole is a key feature of the construct. Although total tests scores are commonly used, the cross-curricular tests developed by ACER can also be partitioned into two sub-tests of MST and HASS.

The ACER cross-curricular construct recognises similarities in mathematical, scientific and technological studies that contrast in some respects with the similarities to be seen in studies of language, arts and of society and the environment. The overall total is meaningful in a cross-curricular test, and the total score can be partitioned into two or more useful sub-scores.

The distinctive characteristics of a cross-curricular test

A cross-curricular test should be distinguished from psychometric constructs like verbal ability and quantitative ability. The stimulus-based and real-world nature of cross-curricular tests also contrasts markedly with tests of abstract reasoning ability. A cross-curricular test aims to be concrete in that candidates are asked to think about something in a meaningful context.

In one sense a cross-curricular test is related to the truly generic skills of reading and writing. As can be seen in the CSSE objectives above, reading and writing were primary aspects of a cross-curricular test, and numeracy (in terms of quantitative reasoning, as distinct from mathematics) was added to the language skills. While they are concerned with the most basic literacy skills, there is an emphasis on reasoning skills in cross-curricular tests that distinguishes them from most literacy and numeracy tests.

The literacy skills of reading and writing (along with reasoning) are the central generic cognitive skills. The generic skills of reading and writing can be applied to or exercised on any content or

across any domain. While numeracy can be treated in a more or less content-specific fashion, numeracy is usually manifest as a kind of thinking in cross-curricular tests.

Figure 1 presents a diagram that typifies the generic and cross-curricular skills test construct that has been used by ACER since the 1960s. The diagram envisages that items can be written to test different areas of the whole model. The outer ring of the diagram is made up of more specialised abilities. The inner ring is differentiated in curriculum or domain-related terms. The core is envisaged as general cognitive ability (generic cognitive skills) and the generic aspects of language skills.

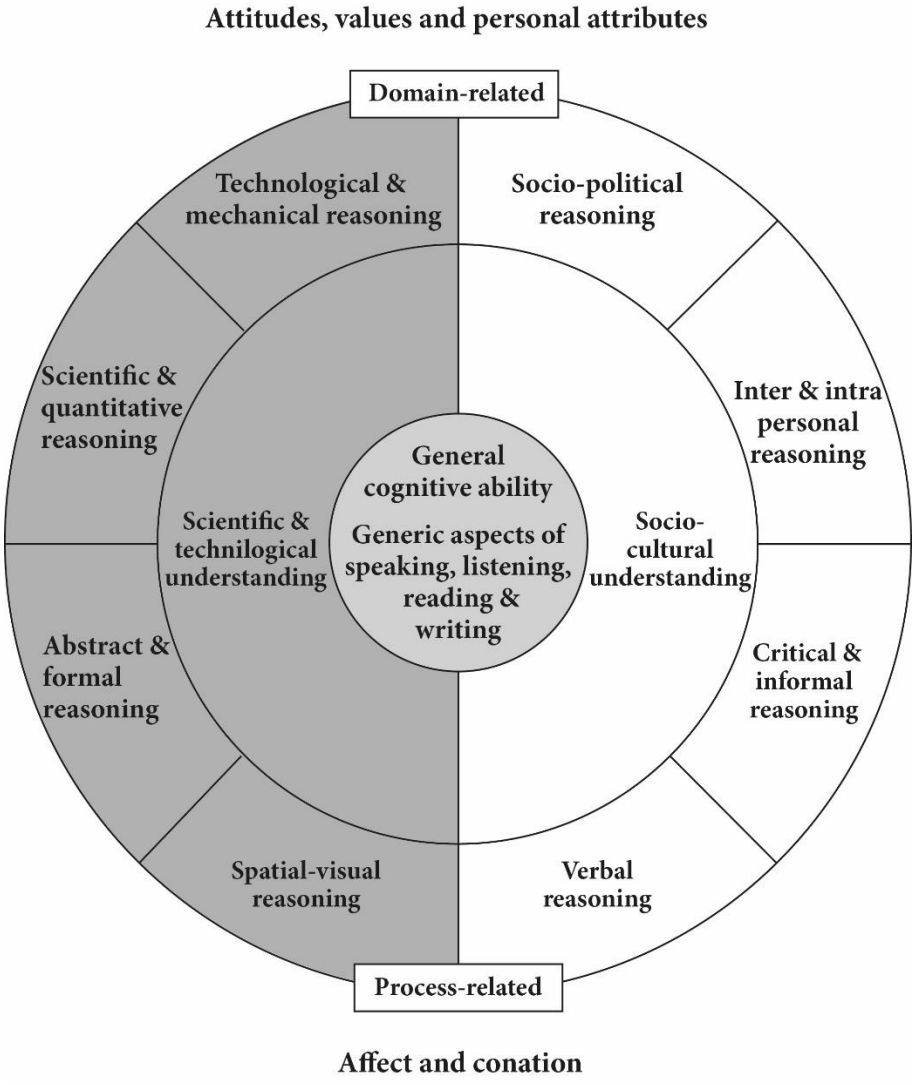


Figure 1. A generic skills model (Source: McCurry, 2005)

The ACER definitions of HASS and MST

The cross-curricular tests currently developed by ACER are constructed on the basis of two sub-tests: **Reasoning in Humanities, Arts and Social Sciences** (HASS) and **Reasoning in Mathematics, Science and Technology** (MST).

While a cross-curricular test assesses generic academic skills, the construct also recognises the way thinking is conditioned by the topic being considered, and the way the topic is represented, as

well as the kind of thinking required by a task or question. Cross-curricular tests are structured and audited in terms of the:

- kind of thinking or thinking processes;
- mode of representing information and ideas; and
- context or topic for thinking.

What follows, are descriptions of HASS and MST in terms of processes, mode and context. While these descriptions are the basis of matrixes that can be used to audit breadth and balance in cross-curricular sub-tests, the fundamental aim of a cross-curricular test is to be a broad and balanced test of thinking skills as a whole.

Reasoning in the Humanities, Arts and Social Sciences

HASS reasoning involves:

- reading and interpreting written text, pictures, diagrams and other data;
- analysing and understanding concepts and the relationship between different concepts;
- grasping the implications of propositions and ideas;
- seeing the direction and upshot of ideas and arguments; and
- identifying the way different ideas and arguments relate to each other.

Different kinds of text are used as stimulus, including passages of personal, imaginative, expository and argumentative writing in a HASS sub-test. Although much of the stimulus material in this section of a cross-curricular test is in the form of written passages, some units may present ideas and information in pictorial, diagrammatic or tabular form. Materials deal with a range of academic and current affairs issues of socio-cultural, political, ethical, personal and interpersonal significance. The material used in the test is usually authentic, existing material rather than being written for the purpose of asking test questions. The questions asked are directly related to authentic interpretation of real-world information presented in a variety of forms.

HASS questions demand varying degrees of

- complex verbal processing and conceptual thinking,
- logico-deductive and plausible reasoning, and
- objective and subjective thinking.

Questions that emphasise verbal processing and conceptual thinking require candidates to recognise explicit and implicit meanings through close reading of words and phrases, and through global interpretations of texts.

While in some sense all HASS questions involve logical reasoning, candidates are also required in some cases to draw inferences and arrive at interpretations that are not irrefragable from a logico-deductive standpoint or objectively or empirically true. These plausible reasoning questions require candidates to distinguish cogent inferences and interpretations from weak ones. Such questions deal in some sense with subjective matters and value judgments, and in answering these questions, candidates have to distinguish plausible interpretations from weak or implausible ones. Questions that emphasise plausible reasoning involve interrelating, elaborating and extending concepts and ideas, and drawing conclusions. Such questions, which require candidates to make discriminations and judgments in the realm of plausible reasoning, test higher-order, critical thinking.

A HASS sub-test aims to be:

- both reasoning and reading;

- interpretation rather than comprehension;
- inferential rather than literal;
- conceptual rather that verbal;
- global rather than particular;
- informal rather than formal reasoning;
- plausible rather than logico-deductive reasoning;
- complex rather than complicated; and
- authentic and meaningful thinking rather than test exercises.

The HASS thinking processes

HASS thinking can be described as having three aspects.

Critical reading: In the area of critical reading HASS items assess the ability to comprehend and interpret written text. Critical reading involves the close examination of particular words, phrases and sentences. It also involves the global comprehension, analysis and interpretation of written texts.

Conceptual reasoning: In the area of conceptual reasoning items assess the ability to reason about HASS related concepts and the relationship between such concepts. Such reasoning will tend to be about less verbally complex material or more formally presented data. As with the critical reading and reasoning processes, conceptual reasoning will tend to be informal and plausible rather than formally logico-deductive reasoning.

Critical reasoning: In the area of critical reasoning HASS items assess the ability to comprehend and analyse contentious issues and arguments about socio-cultural issues. Critical reasoning involves the ability to recognise the way arguments and evidence relate to conclusions, analyse and assess dialectical strategies, and evaluate the strengths and weakness of arguments.

The HASS modes of representation

HASS material is audited in terms of the degree to which it involves the interpretation of more or less demanding written text (emphasising reading or reasoning), and the degree to which it involves interpretation of concept maps, pictures, diagrams, graphs and other data.

HASS contexts

A HASS sub-test is audited in terms of the degree to which it involves humanistic, artistic and social science topics. The *Humanities* are understood to be concerned with the study of socio-cultural ideas. The *Arts* are understood to be concerned with personal and cultural expression and meanings. The *Social Sciences* are understood to involve the study of the socio-cultural world in a systematic and formally theorised fashion. These different aspects of HASS content are interpreted as emphases rather than formal, academic disciplines. HASS items are usually based on general, public discourse rather than more formal academic discourse.

Reasoning in Mathematics, Science and Technology

MST reasoning involves:

- determining facts and objective ‘truth’;
- rigorous logic and justifiable conjecture;
- scientific reasoning (including scientific method);

- quantitative, symbolic, mechanistic, and abstract reasoning;
- analysis, comprehension, application and synthesis of general scientific and quantitative information;
- identification and application of rules and relationship, and determination of consequences;
- planning and organising for a purpose;
- identifying, defining and solving problems;
- evaluating proposed problem solutions;
- critical thinking, justifiable inference, and evaluation of evidence and argument;
- decision making based on data and evidence;
- hypothesis generation based on observation/data;
- hypothesis and inference evaluation; and
- understanding and applying technology.

MST items focus on general logico-quantitative reasoning, which, in the cross-curricular field, is analogous to the logico-mathematical intelligence of Howard Gardner's multiple intelligences theory (1993). It refers to precise critical and logical reasoning that can be applied to quantitative and non-quantitative material alike.

The emphasis of the MST component is on generic skills that are likely to have cross-curricular relevance for typical MST subjects, and some others. Typically such subjects require information analysis and application, problem solving and decision-making based on critical reasoning, with a focus on objective reasoning to determine facts and solutions, and make valid inferences and evaluations.

A basic level of numeracy is assumed in MST assessments. The level varies with test purpose and target population. The term 'quantitative' is sometimes used for the MST component, but in practice, such items focus on critical and logical reasoning based on non-quantitative science-related text or data rather than mathematical knowledge.

MST modes of representation include text, numbers, symbols, diagrams, graphs and tables, with the balance related to test specifications.

MST contexts are generally authentic and accessible but may be classified as mathematics, science or technology. Much of the material is best classified as 'general interest' (social science, non-specialist science, everyday problems). Abstract contexts may be appropriate in certain circumstances.

MST Processes

MST reasoning can be described as having two major aspects:

Apply/Solve

- In the area of *Application* items assess procedural skills such as the ability to apply logical and quantitative skills (including basic calculation, estimation, interpolation, extrapolation), and rules, algorithms, equations and codes).
- In the area of *Problem Solving* items focus on both deliberative and procedural skills (often non-routine), where students have to solve a problem by analysing it to identify its salient features, identifying key information, translating, transforming and synthesizing information into suitable forms, generating a solution strategy and applying it, or evaluating potential solution strategies.

Infer/Decide

- In the area of *Inference* items require students to analyse, extract and interpret information, comprehend or visualise concepts, identify similarities and differences, identify connections, relationships, trends and patterns, identify hypotheses and generalisations, and draw straightforward conclusions.
- In the area of *Decision Making* items require students to go beyond the information presented to generate conjectures, conclusions, generalisations, hypotheses, classifications, syntheses, and assessments and evaluations (e.g., value of evidence, likelihood of an event, usefulness of a strategy, suitability of an experimental design, value of a technological solution).

MST modes of representation

MST items use numerical, graphical, textual, diagrammatic, tabular and symbolic representations to varying degrees. Their emphasis on vocabulary and cultural knowledge deliberately tends to be minimal, though some units focussing on scientific critical reasoning may be based totally on text.

MST contexts

MST items use contexts that allow the application of the processes and modes described above. Though some units are based on contexts that are explicitly scientific or mathematical (in a classroom sense), others use general contexts to which the processes and modes can be applied (for example, social science, non-specialist science, everyday problems). In general, contexts are real-world, authentic and accessible. However, some are abstract or hypothetical (as is often the case with mathematical and scientific work), though still accessible to the students doing the test.

The cross-curricular test matrixes

Test matrixes can be used to check test breadth and balance of cross-curricular tests. The following matrixes, presented in Table 1 and Table 2, map the different aspects that can be used to construct the HASS and the MST components of ACER's cross-curricular tests. Each involves thinking processes, modes of representing ideas and issues, and contexts or topics to be thought about. Each aspect can be assessed on a 0-3 point scale to audit the emphasis given to that aspect in a particular test form. The bottom row of the tables give the target percentage for each facet of each aspect.

Table 1. The HASS Matrix

HASS dimensions for auditing breadth and balance								
Item	**Process** (Rate 0 – 3)			**Mode** (Rate 0 – 3)		**Context** (Rate 0 – 3)		
	Critical Reading	Conceptual reasoning	Critical reasoning	Text load	Visual load	Hum	Arts	Social science
1								
2 etc.								
Target %	40	40	20	70	30	45	20	35

Table 2. The MST Matrix

MST dimensions for auditing breadth and balance								
Item	**Process** (Rate 0 – 3)		**Mode** (Rate 0 – 3)			**Context** (Rate 0 – 3)		
	Apply/ Solve	Infer/ Decide	Verbal /textual	Quantitative /symbolic	Diagram /visual	Math	Science	Tech
1								
2 etc.								
Target %	50	50	33	33	33	40	40	20

The science unit presented in Exhibit 7 from 1969 involves some close reading and understanding of diagrams. The unit is exemplary in that the complicated spatial relationships and patterns have meaning and significance. The best CSSE Science material involves such meaningful and interesting interpretations.

Exhibit 7.
The 'Waggle Dance', Science 1969
Items 35-39

In a number of different species of insects the population consists of several different types of individuals, each adapted for some particular function.

The honey bee colony consists of a single queen, a few hundred drones (males) and thousands of workers (sterile females). Young adult workers act as nurses to feed the larvae which emerge from the eggs laid by the queen. Older workers act as guards at the entrance of the hive, they receive and store nectar, secrete wax for new cells and keep the hive clean. The oldest workers fly from the hive and search for water, pollen and nectar.

The passage below and the diagrams refer to the observations made by an Austrian scientist who found that bees can communicate with each other to describe the location, and distance from the hive, of a food supply.

A scout that has located food, returns to the hive loaded with pollen and performs the 'waggle dance' on the wall of the honeycomb (the dance is called a 'waggle dance' because the bee's abdomen waggles during the performance of the dance). Two distinct types of waggle dance have been identified. If the food is less than 10 metres away the pattern of the scout's dance consists of small circles made first to the left and then to the right. However, if the food is 300 metres, or more, away the dance includes a short 'straight' series of steps between each semi-circle made to the left and to the right. The angles these 'straight' runs make with the vertical describe the position of the food relative to the sun. For instance if the food is in the direction of the sun from the hive, the straight run is vertically upward on the side of the comb. The bee then semi-circles around and repeats the straight run. If the food is in the direction away from the sun, the straight run is vertically downward. If the food is located 60 degrees to the left of the sun from the hive, the straight run is 60 degrees to the left of the vertical, and so on, as illustrated in the following diagrams.

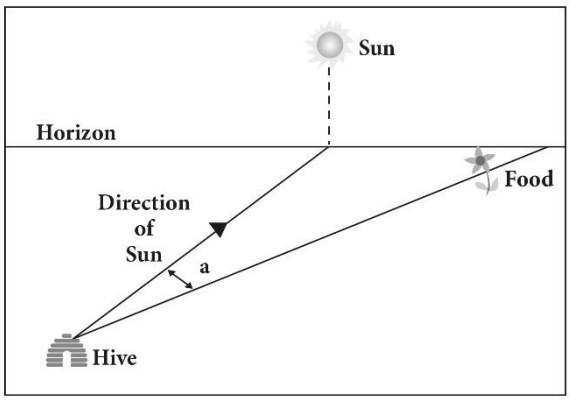


Figure 1

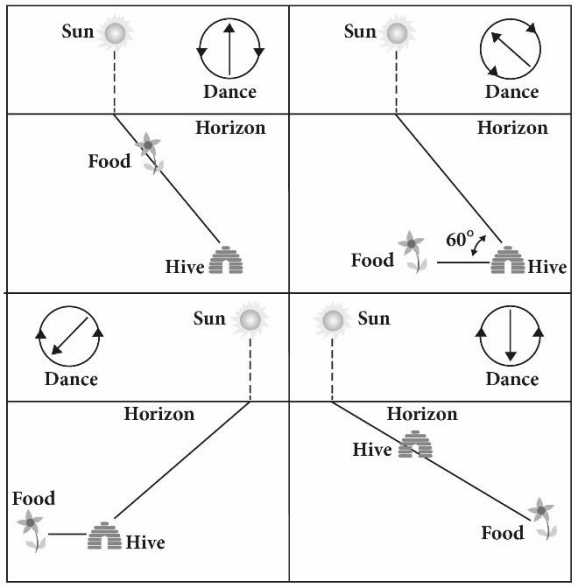
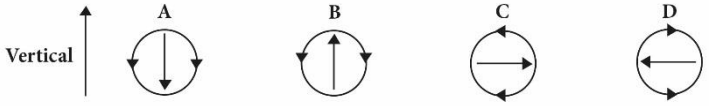


Figure 2

35. The bee scout is
A. a young worker.
B. one of the oldest males.
C. a sterile female.
D. a mature drone.

36. The waggle dance is performed
A. on the ground.
B. in the hive.
C. at the food source.
D. between the food source and the hive.

37. Which one of the figures *A* to *D* below could represent a waggle dance of a scout when the food source, the hive and the sun were in the same relative positions as those shown in Figure 3?



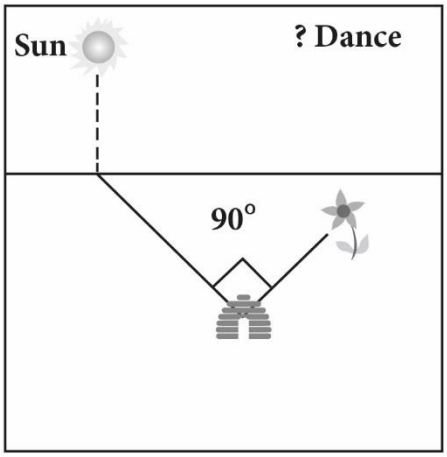
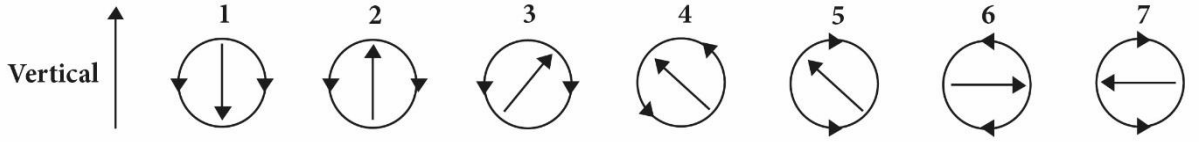


Figure 3

38. The diagrams 1-7 represent waggle dances, two of which are *not* typical dance patterns.



From the following alternatives (*A, B, C, D*) choose the one which lists the two which are not typical dance patterns.

A. 1, 4. B. 4, 5. C. 2, 3. D. 6, 7.

39. Here are five conclusions which might be drawn from the pattern of a waggle dance.

1. The food source is at a distance greater than 300 metres from the hive.
2. The food source is in the direction of the sun from the hive.
3. The food source is more than 10 metres from the hive.
4. The food source is directly above the face of the honeycomb.
5. The waggle dance is performed at least 50 metres from the food source.

Figure 4 shows the stages of a particular waggle dance.

Figure 4

Which two of the above conclusions are correct for this particular waggle dance? Select the correct pair of conclusions from the following alternatives.

A. 1, 2 B. 1, 3 C. 2, 3 D. 3, 5

Exhibit 8 also involves interesting thought about the different ways in which numbers can be used to determine a result. The calculations are simple but how the calculations work is quite complicated.

Exhibit 8.
The Olympic medal tally, Quantitative thinking 1970
Items 18-22

The following table shows the number of gold, silver and bronze medals won by each of thirteen countries at the 1968 Olympic Games. In reporting the results of the Games, newspapers ranked the countries on the basis of the number of gold medals won. Where two or more countries won the same number of gold medals, the number of silver medals won was used as the basis for rating. The table below shows the first thirteen place-winners in this unofficial contest.

		Gold	Silver	Bronze	Total
1	USA	45	28	34	107
2	USSR	29	32	30	91
3	Japan	11	7	7	25
4	Hungary	10	10	12	32
5	E. Germany	9	9	7	25
6	France	7	3	5	15
7	Czechoslovakia	7	2	4	13
8	W. Germany	5	10	10	25
9	Australia	5	7	5	17
10	Britain	5	5	3	13
11	Poland	5	2	11	18
12	Rumania	4	6	5	15
13	Italy	3	4	9	16

If the countries were ranked on the basis of the total number of medals won, they would not appear in the same order.

18. Which of the following countries would rise most in the rankings if places were given on the basis of the total number of medals won?

A. Hungary B. Australia C. Britain D. Poland

19. Which of the following countries would have the same rank on the basis of total medals won as it has in the table?

A. Japan B. Australia C. Rumania D. none of the above

Questions 20 and 21 refer to the following additional Information:
Some newspapers reported unofficial results of the Olympics on the basis of a points system, i.e. x points for a gold medal, y points for a silver medal, and z points for a bronze medal. In each case x > y.> z.

20. Which of the following sets of values of x, y and z results in Japan being placed ahead of Hungary?

A. 10, 6, 2 B. 12, 8, 4 C. 15, 5, 1 D. 20, 4, 1

21. In which of the following pairs is the order incapable of being reversed, whatever the values assigned to x, y and z?

A. France-Australia B. Hungary-Japan C. France-Czechoslovakia D. Britain-Rumania

22. If Italy finished with a total score of 32 under a system of scoring whereby bronze medals are worth one point each, how many points are awarded for gold and silver medals respectively?

A. 4, 3
B. 5, 2
C. It is possible to find a simple answer but it is not A or B above.
D. It is not possible to find a simple answer.

4

The development of cognitive testing

To fully understand the distinctive features of the CSSE/ASAT style of testing described in Chapter 3, these tests need to be compared with the predominant style of testing that developed within cognitive psychology and the discourse of educational measurement during the 20th century.

The CSSE was developed after some 50 years of cognitive testing, and some 30 years after the foundation of ACER. In procedural terms, the CSSE and ASAT are psychometric tests in that they are developed through a trial testing procedure, and the statistical properties of test forms are controlled and monitored in a psychometric fashion. The CSSE/ASAT were developed using similar procedures as more typical psychometric tests like the SAT (once but not now known as the Scholastic Aptitude Test) of the College Entrance Examination Board in the United States. But as discussed later, the CSSE/ASAT contrast significantly in texture and tenor with the SAT.

This chapter sketches the major lines of development in cognitive psychology and educational measurement over the last century, with particular attention to what is called the differences between the Spearmanian generalists and the Thurstonian specialists. In the process, the distinctive characteristics of the CSSE/ASAT style of testing will become more evident.

The major approaches to understanding abilities in cognitive psychology

Attempting to understand the human mind has been at the centre of the scientific study of psychology that has developed over the last 120 years. A key aspect of this development has been the attempt to understand human learning and intellectual performance, which has in turn prompted thinking about the nature of cognitive ability, and different kinds of cognitive ability.

Psychology, and particularly the related branches of cognitive psychology and psychometrics, have been the areas of systematic inquiry that explore notions of cognitive performance and notions of generic cognitive abilities. Cognitive psychology and psychometrics, in particular, have developed various differential ability theories.

This psychology of cognition has its roots in phrenology, and phrenology was a primitive kind of differential ability theory (Fodor, 1983). While the CSSE/ASAT tradition is not explicitly conceptualised in relation to the psychometric approach to cognitive testing, the CSSE/ASAT tradition needs to be compared with typical psychometric tests to be seen for what it is.

Two major streams of development might be identified in the history of cognitive psychology over the last century. One stream has attempted to understand the mechanisms of thought, and this stream has contained both psycho-biology (or wet psychology) and another stream of more experimental and or theoretical psychology. This experimental and or theoretical stream developed into cognitive psychology and the more general field of cognitive science and artificial intelligence in the last decades of the 20th century (Gardner, 1985). Another stream of development has paid particular attention to individual differences in intellectual performance, particularly in the sub-branch of psychology called psychometrics (Carroll, 1982; Carroll, 1993).

Psychometrics is the application of techniques of statistical analysis to cognitive performance, particularly the performance of groups of candidates in formal testing situations. As will be seen below, the psychometric approach involved the development of techniques for the statistical analysis of test performance. Ironically the psychometric approach is sometimes referred to as the psychology of individual differences, not because it deals with individuals but because it deals with the differences within groups.

The psychometric approach along with the behaviourist paradigm came to dominate psychology, particularly in the United States, in the first half of the 20th century. The 1950s and 60s saw increasingly widespread dissatisfaction with the reluctance of behaviourism to countenance or deal with notions of mental processes. There developed a marked desire in this period to analyse cognitive processes, particularly through the use of computers to model thinking. These developments in the 1970s came to be known as 'the cognitive revolution' (Gardner, 1985).

The various approaches to understanding cognition that developed with the 'cognitive revolution' contrasted with the psychometric approach either because they were concerned with what is general in cognition (as in the work of Jean Piaget or artificial intelligence research) and were not concerned with 'individual differences', or because they aimed to analyse and understand thought processes in general. The differences between the psychometric and the cognitivist approaches to cognition may be crudely summarised as follows in Table 3 (adapted from Ceci, 1990).

Table 3. The differences between psychometric and cognitivists approaches to ability

Psychometric		Cognitivist
Empiricism	⟷	Theory
Operationalist	⟷	Analysis
Experimental	⟷	Rational
Statistical data	⟷	Eclectic data
Instrument building	⟷	Model building

The major approaches to understanding cognition

A survey of the last 100 years of theoretical and empirical research into human abilities suggests a number of contrasting approaches and lines of development that might be named as follows:

- psychometric or operationalist theories,
- cognitive-correlate or bottom-up theories,
- information processing or componential theories,
- psycho-biological theories,
- cognitive development theories, and
- socio-cultural or contextualist theories.

The psychometric or operationalist theories focus on the differences between individuals in performance and observable outcomes. In differing degrees this approach is influenced by the behaviourist reluctance to contemplate mental processes in favour of the 'observable behaviour' of test performance. In the second part of the 20th century psychometrics became more concerned with attempting to discern and theorise cognitive processes in test performance.

The psychometric approach is that taken by Spearman, Thurstone, Vernon, Cattell and Carroll, as outlined below. In his review of the work in this tradition, Spearritt (1996, p.112) describes this approach as one in which "the identification of a cognitive ability is based on measures of correlation or other forms of association among cognitive tasks."

Another approach that developed out of the psychometric approach as part of the cognitive revolution of the 1970s came to be known as the 'cognitive correlate theories'. This work sought to find significant and meaningful correlations between psychometric indicators and performances on various elementary cognitive tasks. This approach can be seen in the work of Pellegrino and Glaser (1979) in which they sought to specify the intellectual abilities that are differentially related to high and low levels of scholastic aptitude. Jensen and Munro (1979) have sought to test the basic processes like the time taken to react to a stimulus (responses that are not seen as matters of intelligence or knowledge) and to then correlate the results of such experiments with intelligence test scores. These bottom-up theories seek some form of speed of neural processing or nerve conduction as explanations of differences in cognitive ability. The cognitive-correlates approach is represented by the work of Earl Hunt and his colleagues (1975) who asked: "What does it mean to be high verbal?" and then sought to determine how basic verbal processes like lexical retrieval relate to overall verbal performance on tests of verbal ability. The cognitive-correlates approaches are termed 'bottom-up' in contrast with the 'top-down' approaches of Binet and Vygotsky mentioned below.

The development of theories of information processing in the 1950s and 1960s prompted proponents of the cognitive components approach to ask the question: 'What does the intelligence test, test?' This approach aimed at both theoretical and empirical analyses of test performance with the intention of building precise theories of knowledge and cognitive processes. Such work would typically begin with 'task analysis' of test items so as to identify information processing components, then develop a model of task performance, and then apply such models to research into the differences in the performance of individuals.

The Yale psychologist, Robert Sternberg, was the major proponent of the information processing approach, and he began his eclectic cognitive research within this paradigm. In his PhD research Sternberg sought to decompose analogical reasoning tasks from conventional psychometric tests like the SAT into separate activities and then relate the separate component activities to overall intelligence (Sternberg, 1977). This approach can be seen to have informed Sternberg's influential triarchic theory of intelligence.

Psycho-biological theories of cognition focus on the relationship between the brain and the mind. They try to bridge the gap between psychology and neurology by attempting to link basic activities of the brain with complex frontal-lobe activities. Howard Gardner's multiple intelligences theory is in part a version of this approach (Gardner, 1985). Linking the bottom-up and top-down theories or the psycho-biological and cognitive theories is the overall, but still distant, goal of cognitive research.

The cognitive development perspective focuses on the general or typical path of cognitive development from birth to maturity. This approach seeks to trace the developmental trajectory of the growing child in an attempt to understand human capacities in general. It seeks the general theoretical principles that govern intellectual growth. Jean Piaget is the initiator and dominant

representative of this approach (Piaget, 1971). Howard Gardner was significantly influenced by this approach, and he seeks to accommodate aspects of Piaget's work into his multiple intelligences theory.

Socio-cultural or contextualist theories of cognitive performance focus on the relationship between cognition and cultural context. The Russian psychologist, Lev Vygotsky, came to be seen as a major influence on these theories (Vygotsky, 1986). In contrast with Piaget, Vygotsky saw the development of cognitive abilities as a matter of internalising cultural practices rather than a process of the individual accommodating to the world. For Vygotsky learning and intellectual development are largely shaped by a child's interactions with others. These theories look past individuals to the way a social context shapes intellectual performance, and they give particular attention to the way culture shapes the development of cognition. Such approaches emphasise the way intelligent behaviour differs in different cultural contexts. These approaches are relativistic rather than hereditarian or nativist. They see intelligent behaviour as socially situated and culturally mediated. The bio-ecological theory of Stephan Ceci is a major representative of this approach (Ceci, 1990).

As was mentioned above, the psychometric approach to understanding cognitive abilities was predominant in the first two thirds of the 20th century, and this approach then evolved into the cognitive correlate and cognitive component approaches. Developmental theories were prominent in Europe, but they had little impact on cognitive theory in the English speaking world until the 1970s. The influence of Piaget and the discovery (re-discovery?) of the work of Vygotsky came to have an impact on cognitive theory in the 1970s as the influence of behaviourism waned in the English speaking world. These developments challenged the position of psychometric theories of cognition in the 1970s, and in the last 25 years of the 20th century, developmental, socio-cultural and psycho-biological theories of cognition came to challenge and rival psychometric theory and psychometric practice.

Criticism of the psychometric approach

Criticism of the psychometric approach has become more common and more vociferous in the last 30 years. The psychometric tradition has been accused of being atheoretical or theoretically weak, and developmental, socio-cultural and psycho-biological theories have gained increasing attention in recent times. Although there is an increasing acceptance of so-called 'performance assessment' in the United States (the place where psychometric testing is most powerful), the tests of the psychometric tradition are still deeply entrenched in that system. These tests also provoke considerable opposition from some educators and are matters of controversy in the popular media.

The psychometric approach to understanding abilities has been criticised for being:

- based on narrow pencil and paper tasks;
- based on flawed and inconclusive statistical methods;
- merely operational and atheoretical;
- without substantive, real world meaning and predictive power; and
- culturally biased and hence inequitable.

The following discussion outlines the development of the psychometric tradition in the 20th century and the way it was challenged by the other approaches mentioned previously. The strengths and weaknesses of that approach are reviewed, and the style and tenor of psychometric testing is compared with the CSSE/ASAT approach in the next chapter.

Psychometric theories of ability

Francis Galton and the positive manifold

The history of psychometric theory and practice begins with Frances Galton (1822-1911), the second cousin of Charles Darwin. Galton was one to the progenitors of scientific psychology, and he had a particular interest in kinds of cognitive ability and differences in such abilities. For Galton, the scientific study of psychology was based on statistical evidence and proof. This commitment to the statistical exploration of test performance is the basis of the psychometric exploration of ability.

Galton can be described as taking a bottom-up or cognitive-correlates approach to understanding complex cognition. He was particularly interested in different sensory and motor functions and how they related to complex intellectual performances. He studied a variety of psycho-physical tasks, such as weight discrimination and sensitivity to musical pitch, in an attempt to understand complex intellectual performance (Forrest, 1974).

Other researchers followed the example of Galton in exploring such psycho-physical tasks as a means of predicting intellectual ability from basic processes, but by the end of the second decade of the 20th century it seemed that little or no relationship between higher cognitive performance and basic processes could be demonstrated.

Galton also worked on complex cognition by comparing the performance of groups of students in different school subjects. To undertake such studies, Galton formulated the correlation coefficient statistic for making comparisons of the performance of students in different subjects, and hence began the use of statistical correlation (the basic technique of the psychometric approach) for the study of intellectual performance. These techniques were extended and refined by Galton's student and co-worker, Karl Pearson, and the process of statistically analysing cognitive performance that became the psychometric approach had begun.

As well as some striking ideas about cognition, some of the more distasteful aspects of psychometrics can be traced to Galton. It seemed to Galton that genius ran in families and intellectual ability was substantially inherited. He was a convinced hereditarian and believed that a significant amount of an individual's intelligence was genetically determined. He coined the word 'eugenics' and was a major proponent of the eugenics movement that advocated improving the human species by selective breeding (Gould, 1996).

Galton's correlational studies of the performance of students in different subjects revealed high positive correlations (Galton, 1883). This fact (or artefact to some) came to be known as the 'positive manifold'. The positive manifold is a significant issue in the theory of cognitive ability.

The work of Galton initiated the long-running, and perhaps eternal debate, between the advocates of a general cognitive ability and those advocating separate, specific abilities. Galton discovered that there was a positive correlation between the school grades of pupils in quite different subject areas. The pupils he studied showed a general pattern of strength or weakness across different subjects, and this prompted him to ask whether there was one or a group of abilities that were involved in different subject areas.

The positive manifold of cognitive performances was found to hold true in almost all instances (on reasonably broad samples of the population) and these observations gave rise to a theory of general abilities or even one general ability that underpins all cognitive performances. On the other hand, it has been argued that the differences between different kinds of cognitive performance are very substantial and that these differences are as important and as informative as the similarities. This second argument has emphasised different special abilities.

The meaning and the significance of the positive manifold and the theory that grows from it has been the subject of on-going dispute. The extent to which a cognitive theory recognises the positive manifold as significant is the key difference between most theories. Table 3 above, adapted from the work of Ceci (1990), compares the different approaches to intelligence. The positive manifold was used as the basis of an argument by some for the existence of an overall or general ability or intelligence. Instruments were developed to test this general intelligence and the statistical technique of factor analysis was used by both sides of the debate to support their advocacy of the pre-eminence of either general or special abilities.

Alfred Binet and intelligence

The second significant development in the understanding of cognitive abilities took place when the French physician, Albert Binet (1857-1911), was requested to develop a method for testing the intellectual abilities of children after the introduction of compulsory education in France. The purpose of the testing was to determine which children lacked the ability to cope with normal schooling and should be placed in special schools.

Binet and his co-worker Simon rejected the testing of sensorimotor functions explored by Galton and focussed instead on higher cognitive functions (Binet & Simon, 1916). In contrast to Galton's bottom-up approach, Binet aimed to assess complex mental functions from the top-down. For Binet judgement was the basis of intelligence. He identified three major elements in intelligent thought: direction; adaptation; and criticism. But Binet's approach was pragmatic in that he sought to discover kinds of task that would differentiate between levels of development without trying to develop a theory as the basis of this differentiation.

Binet and Simon created their first intelligence scale in 1904 to discriminate between normal and mentally deficient children. In 1908 a further version of this scale was developed for differentiating among normal children. Binet's intelligence test assessed skills such as judgment, comprehension, and reasoning, and he sought to use sequences of tasks to define the usual pattern of growth. Unlike the work of Galton, the focus of Binet's work was on more complex mental processes, and the tests developed by Binet, unlike the tests of Galton, seemed to successfully identify significant differences between the abilities of individuals. Binet's tests were very successful and they were a major stimulus to the development of standardised testing and the psychometric approach to understanding ability.

Lewis Terman (1877-1956) of Stanford University produced an American version of Binet's tests and he also developed the intelligence quotient (the IQ) as a standardised method of representing relative intellectual ability within and across age groups. The Stanford Binet (as Terman's version of the test was and still is known) began the remarkable development of mass, standardised testing in the United States. The Stanford Binet has undergone some adaptation over the last 90 years, but it remains clearly related to the tests originally developed by Binet.

It seems that the psychometric approach was developed from a synthesis of the statistical methodology of Galton and the test materials of Binet.

Charles Spearman and factor analysis

The British psychologist Charles Spearman (1863-1945) was particularly inspired by Galton's correlational work, and in using the tools developed by Galton he also found that seemingly diverse tests were strongly correlated (Spearman, 1923; Spearman, 1927):

> *The earliest fundamental observation made was that the inter-test correlation, although widely varying in magnitude, were at least regularly positive in sign.*

Although there was a high correlation between the cognitive tests used by Spearman, there was also a clear pattern or clustering of higher and lower correlations. It seemed to Spearman that there was a kind of hierarchy in the correlations between different tests. In an attempt to reduce the complexity of the correlations between different tests, Spearman made the first steps in the development of the statistical technique called factor analysis that is at the heart of the psychometric tradition.

The development of factor analysis began with Spearman's use of the correlation coefficient that Galton had designed for the study of the relationship between the performances of individuals on different tests. Factor analysis developed a complex mathematical rationale and took many forms as it developed throughout the 20th century, but its basic aims and characteristics remain the same. In his *Dictionary of Statistics and Methodology*, Vogt (1993) defines factor analysis in the following terms:

> *Any of several methods of analysis that enable researchers to reduce a large number of variables to a smaller number of variables of 'factors' or latent variables. Factor analysis is done by finding patterns among the variations in the values of several variables; a cluster of highly correlated variables is a factor.*

One of the industry standard statistical applications for undertaking factor analysis, the *Statistical Package for the Social Sciences*, describes factor analysis in similar terms:

> *Factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables. Factor analysis is often used in data reduction to identify a small number of factors that explain most of the variance observed in a much larger number of manifest variables (SPSS 9.01).*

Gould (1996, p.67) describes factor analysis as a technique that "simplifies large sets of data by reducing dimensionality and trading some loss of information for the recognition of ordered structure in fewer dimensions". Anderson summarises the purpose of factor analysis as follows (Anderson, 1992):

> *What factor analysis does is to take a data matrix of intercorrelations and search for underlying traits that could have generated the patterns of co-variance seen in the data matrix.*

Factor analysis represents the aim of ability theory to cluster different kinds of tasks and kinds of thinking into an overall framework.

Through the use of factor analysis, Spearman came to believe there was one significant factor that would explain a large part of the individual differences to be seen in a range of different test performances. Spearman came to view intellectual ability as a kind of 'mental force or energy operating through many specific engines' (Spearman, 1946). He defined intelligence as 'eduction of relations and correlates' which involved making inferences about general rules and correlates or the deduction of specific instances from a given general rule.

Spearman developed what is known as the 'two factor theory' which claims that any cognitive performance involves two factors, a universal, general factor that is deployed in any cognitive task and a specific factor for that particular performance. This notion of one large factor of 'general intelligence' or (*g*) and a small number of specific factors (*s*) for particular tasks has been very influential. Spearman (1927, p.88) claimed optimistically that "the frank adoption of two factors would seem to supply an adequate basis – the only one possible – for a unified science of human ability".

On one hand, Spearman side-stepped some of the most fundamental and difficult questions about mental abilities in a way that typified much of the psychometric approach (Spearman, 1946):

> *We are under no obligation to answer such questions as: Whether 'factors' have any 'real existence'? Do they admit of genuine 'measurement'? Does the notion of 'ability' involve at bottom any kind of cause, or power? or is it only intended for the purposes of bare description? How, if at all, does the role of factors transcend that of classes? (p.25)*

Even so, Spearman did not make unguarded claims or insinuations about what had been achieved by the statistical study of test performance:

> *In spite of the attempts to bridge the gap between that statistical study of factors and their psychological interpretation, no one can say that a final interpretation of any broad factor has been obtained. (p.191)*

Although Spearman gave emphasis to the general factor, he also identified broad qualitative factors. He introduced the basic verbal and mechanical factor distinction that was developed in Britain by Cyril Burt (1940) and Philip Vernon (1973).

L. L. Thurstone and primary mental abilities

Like Spearman, the American psychologist, L.L. Thurstone (1887-1955) saw great promise in the statistical analysis of test performance using factor analysis (Thurstone, 1938):

> *One of the oldest psychological problems is to describe and to account for the individual differences in human abilities. How are these abilities and the great variations in human abilities to be comprehended? For centuries philosophers have been free to set up arbitrary classifications of personality types and lists of abilities, and there have been almost as many classifications as there have been writers. The factorial methods have for their object to isolate the primary mental abilities by objective experimental procedures so that it may be a question of fact how many abilities are represented in a set of tasks, and whether a particular objective performance represents an ability that is in some fundamental sense primary*.

Like Spearman, Thurstone looked to the statistical analysis of test performances to shed light on human abilities, but his analysis of test performance led to quite a different conclusions. Thurstone developed a method of multiple factor analysis with which he analysed 56 variables into 11 or 12 uncorrelated primary factors. He claimed that seven or eight of these factors were clearly interpretable and he saw little evidence in his data of the general ability identified by Spearman.

Thurstone identified seven primary mental abilities from his analysis of test performances, presented in Table 4.

Table 4. Thurstone's primary mental abilities

Word fluency	producing words with an initial given letter, suffix or prefix
Verbal comprehension	knowledge of vocabulary and ability to read
Number	numerical computation and arithmetical reasoning problems
Spatial visualisation	mentally manipulating objects in two and three dimensions
Inductive reasoning	identifying rules in word and number series tasks and analogies
Memory	memorising and recalling words and pictures
Perceptual speed	identifying similarities and differences in graphical and numerical material

Thurstone's primary mental abilities was the first comprehensive differential ability construct developed from psychometric analysis, and it is mirrored in more current theories which are outlined below. As a result of his factor analytic work Thurstone produced the Primary Abilities Test of differential abilities which was widely used between the 1930s and the 1960s. Thurstone's

primary abilities influenced the Differential Aptitude Tests, the most influential of the multiple abilities batteries currently in use.

Like most psychometric theories of ability, Thurstone's primary abilities are a rather heterogeneous grouping. The first three are kinds of knowledge (related to the notion of crystallised intelligence outlined below). The fourth and fifth are kinds of manipulative processes or kinds of thinking (related to fluid intelligence and hence contrasted with crystallised intelligence). The second last of Thurstone's abilities is a basic cognitive function, and the last is an outcome or characteristic of the functioning of other processes.

P. E. Vernon and a hierarchical model of intelligence

A major step in the development of cognitive ability theory was taken by British psychologist and factor analyst Phillip Vernon (1905-1987). Vernon was the first to attempt a full scale reconciliation of the positions of Spearman and Thurstone (the generalists and the specialists factorists) by developing a hierarchical view of the structure of abilities, presented in Figure 2.

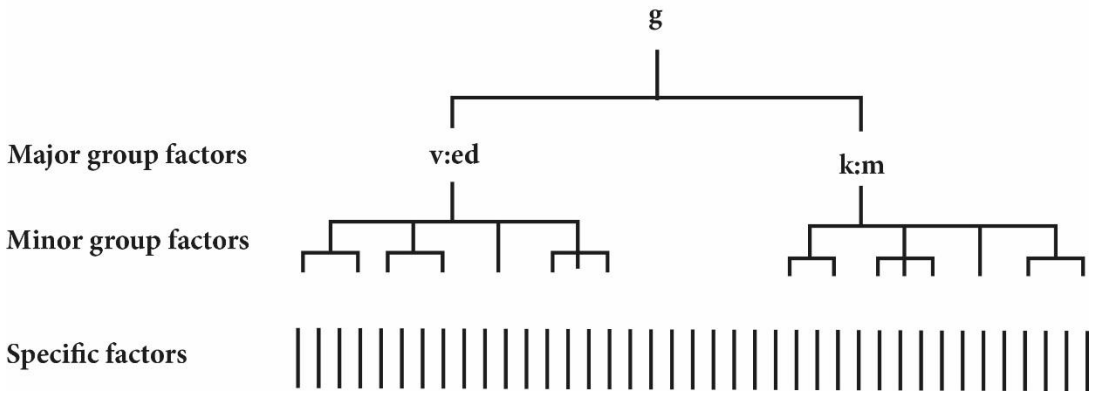


Figure 2. Vernon's hierarchical model of intelligence

In his book *The Structure of Human Abilities*, Vernon (1973) reviewed the results of factor analytic studies to that time. He identified a general intellectual factor and two major group factors:

- *v:ed* or verbal/educational ability, and
- *k:m* or practical/mechanical ability.

He also identified various minor group factors and specific factors that he organised into a hierarchical model that envisaged a general ability which correlated with the specific and the group factors.

Although Vernon recognised the importance of factors other than *g*, he saw *g* as accounting for a significant amount of the differences in test performance. He believed 'most of the variance of human abilities in daily life is due to *g*', and he also believed *g* was more important than specific aptitudes in job success.

Vernon took a rather atheoretical and behaviourist approach to defining and testing cognitive abilities. Where Thurstone used psychometric constructs to develop his theory, Vernon based parts of his work on areas of academic study and skill. The generic skills of reading and handwriting were included in Vernon's model, but so were specific subject areas. Vernon's model of abilities, shown in Figure 3, placed *v:ed* at the centre of a constellation of abilities, and closest to *g*. The work of Cattell moved, on the other hand, away from a notion of *g* at the centre of *v:ed*.

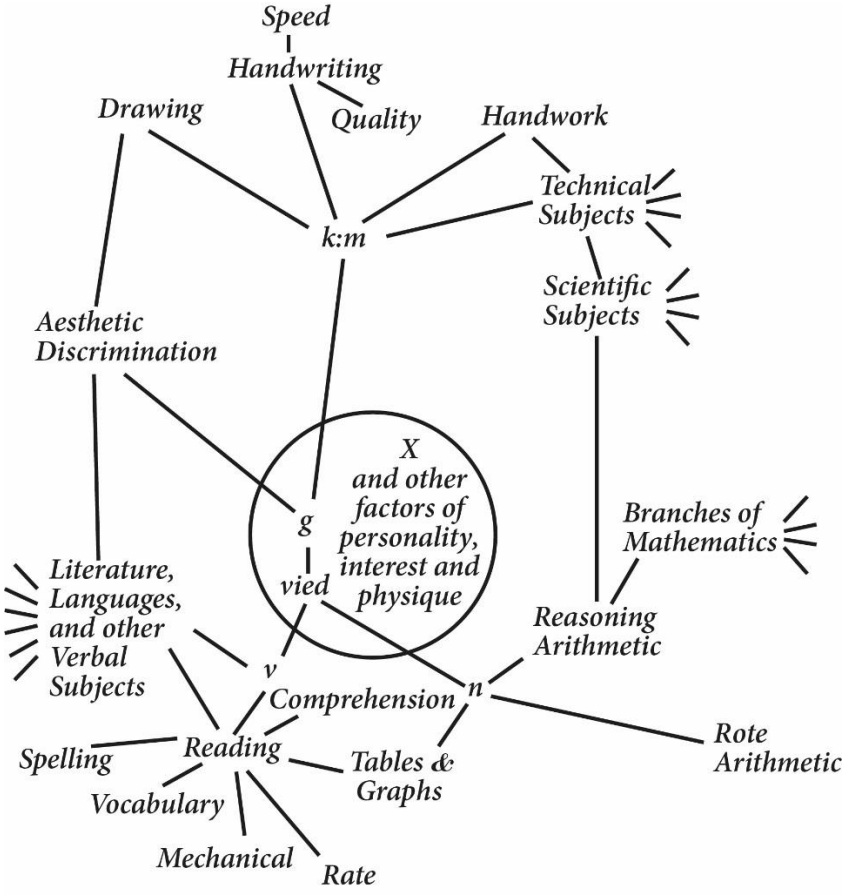


Figure 3. Vernon's structure of educational abilities (p.47)

R. B. Cattell and Gf-Gc theory

A more sophisticated reconciliation of the contrasting views of Spearman and Thurstone was produced by Raymond Cattell (1905-1998) and John Horn in the 1970s (Cattell, 1987). They developed a hierarchical theory (known as *Gf-Gc* theory) which identified two major types of general factors and three minor ones.

Cattell called his position a 'triadic theory' made up of: capacities, provincial powers and agencies. Cattell's theory is essentially a two stratum one including a hierarchical combination of general and special abilities. Cattell's Triadic Theory of Intelligence is presented in Figure 4.

In Cattell's theory, capacities are abilities reflecting limits to brain action as a whole; provincial powers are types of local organisation for different sensory and motor modalities; and agencies are abilities to perform in different areas of cultural content, acquired through the investment of fluid intelligence in learning.

Cattell also made an influential distinction between two kinds of intelligence which he described as 'crystallised' on one hand and 'fluid' on the other. According to Cattell, fluid intelligence is a flexible capacity that deals with novel challenges and crystallised intelligence is the ability that is developed over time as a result of experience. Cattell saw crystallised ability as a result of the 'investment' of fluid ability.

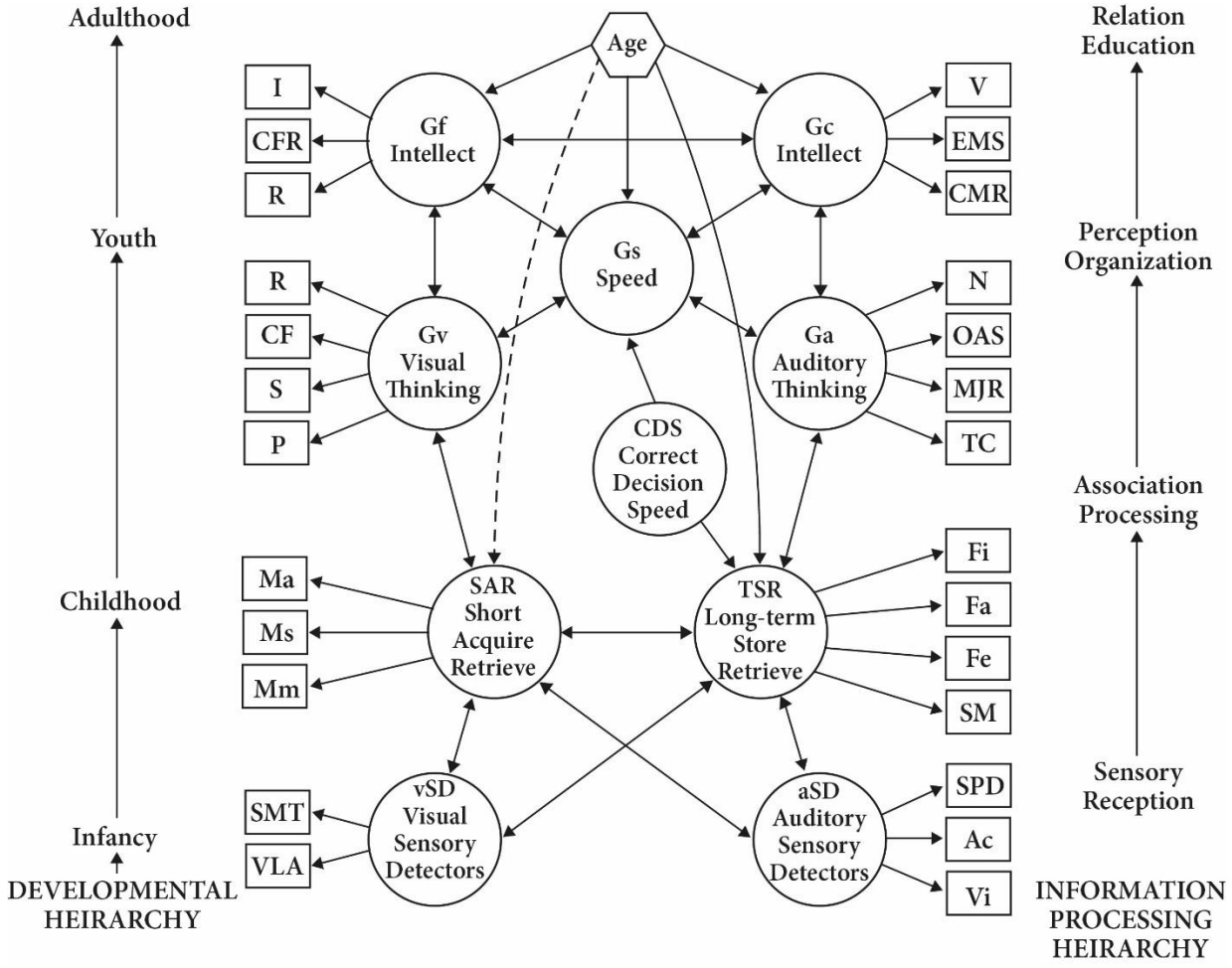


Figure 4. Cattell's Triadic Theory of Intelligence (Encyclopaedia of Intelligence, p.444)

In general terms, Cattell offered the following breakdown of intelligence, presented in Table 5.

Table 5. Cattell's Gf-Gc Theory of Intelligence

Gc	Verbal knowledge, reading comprehension, and prior educational achievement (similar to Verbal-educational factor of Vernon) required for school activities
Gf	Abstract, non-verbal reasoning that is a basic biological capacity that can be measured as perceptual ability (quantitative and other kinds of reasoning relate to this factor)
Gv	Spatial relations, visualisation (related to the practical-mechanical ability of Vernon)
Gr	Memory retrieval
Gs	Performance speed

According to Cattell, verbal tests involved crystallised intelligence and what is commonly meant by 'general intelligence' is best seen in verbal intelligence tests such as the Stanford Binet or the Weschler intelligence tests. On the other hand fluid intelligence is elicited by non-verbal tests such as Cattell's own culture fair test (Cattell, 1987) and the Progressive Matrices developed by Raven to test Spearman's *g* (Raven, 1936). Fluid abilities are the reasoning and problem-solving capacities measured by such psychometric test items as the analogies, classifications, and series

completions. Crystallised abilities can be said to derive from fluid abilities and can be viewed as their products. Tests of crystallised ability include vocabulary, general information, and knowledge about specific fields.

In Cattell's model, the omnibus view of *g* typified by the Stanford Binet or the Weschler intelligence tests has given way to an abstract non-verbal reasoning of the kind found in Raven's Progressive Matrices.

Cattell attempted a comprehensive theory of cognitive performance that included a view of development, a view of basic information processing, a view of higher order abilities, and a view of general abilities. Cattell's hierarchical model has been influential, particularly in the distinction between crystallised and fluid intelligence. According to Carroll, Cattell's analysis of abilities 'comes very close to being a cognitive analysis' (Carroll, 1993).

The cognitive revolution

A quite different approach to the factor analytic tradition was gathering momentum under the influence of multi-disciplinary cognitive science and artificial intelligence (AI) research in the 1970s (Gardner, 1985).

Allen Newell and Herbert Simon developed innovative and influential computer models of cognition, and it seemed that AI was about to shed significant light on intelligent performance (Gardner et al., 1995). A range of researchers sought to use computers to solve problems and to learn. These general problem solving algorithms and heuristics that were built up through AI research had impressive results with chess and mathematical problems. In time, however, the inability of AI machines to reason with informal common-sense or to process natural language made the differences between AI and human thought glaringly evident.

The importance of a broad and rich knowledge base to human cognition was raised to prominence in the late 1970s and 1980s because of the limited success of the generalised algorithms used for problem solving in AI. Although there were impressive developments with neural networks and parallel distributed processing during the 1970s and 1980s, optimism about understanding human cognition through AI receded. By the 1980s it seemed that the information processing approach to higher cognitive functions could not find a model for human thinking in current computers, and the distinctive characteristics of human thought remained elusive (Das et al., 1975).

During the 1970s and 1980s criticism of the psychometric approach led to increased efforts to understand cognitive processes as parts of a coherent explanation of cognition. Under the influence of the Russian theorist Luria, Das and his colleagues (1975) developed a theory of cognitive processes. They tried to relate mental ability to neurological function by developing a notion of simultaneous and successive modes or styles of cognitive processing. This work influenced the intelligence tests produced by Kaufman in the 1980s (Kaufman, 1990).

The cognitive-correlates approach taken by researchers like Earl Hunt (Hunt et al., 1975, Hunt 1978) involved a bottom-up analysis and understanding of cognition. Such researchers looked for correlations between intelligence test scores and basic cognitive activities like speed of reaction to stimulus. Taking the same approach, Arthur Jensen used the most basic tasks that involved little knowledge or information and tried to determine the extent to which processing speed and efficiency are related to intelligence (Jensen, 1998).

J. B. Carroll

The development of the factor analytic tradition peaked with the publication of John Carroll's (1993) *Human Cognitive Abilities: A survey of factor-analytic studies*. In this compendious work,

Carroll undertook a consistent reanalysis of 461 factor analytic data sets, the most important data sets produced between the 1930s and the 1990s. His aim was "to present what is known about cognitive abilities and its scientific basis" (p.28).

For Carroll, the factor analytic approach is a matter of 'empirical studies of ability – a science of human abilities'. He describes his enterprise in the following terms:

> *The investigations dealt with in this book can be regarded as attempts to identify abilities by systematically classifying different tasks with respect to the abilities they appear to require. (p.9)*

and

> *A factor, if it is well established in a number of empirical investigations, is in essence a latent trait reflecting differences over individuals in ability characteristics or potentials. (p.20)*

The factors defining abilities at each of the three levels of his model are observable differences in the performance of individuals on particular classes of tasks.

Carroll saw a fundamental unity in the psychometric attempts to define ability:

> *All models of the structure of cognitive abilities appear to be attempts to classify the various manifestations of mental ability according to aspects of content, type of processing required and type of response or outcome. (p.67)*

As a result of his re-analysis, Carroll produced a three-stratum hierarchical theory according to which cognitive abilities were either narrow, broad or general:

> *There exist a substantial number of distinguishable and important mental abilities – as many as 30 or more. While it may well be the case that general intelligence – a recognised higher-order factor of cognitive abilities – is the most weighty element in all these relationships … the possible importance of more specialised abilities cannot and should not be ignored. (p.27)*

The three levels of Carroll's theory were:

1. the first level of narrow abilities that reflect experience or learning, and are related to Thurstone's primary abilities;
2. the second level of broad abilities which are moderately specialised and differ in terms of process, content and manner of response; and
3. the third level of a general cognitive ability.

Carroll accepted Spearman's notion of *g* and indicated that his results were consistent with Spearman's two factor model. But Carroll did not see fundamental differences between the work of Spearman and Thurstone, and his model incorporates the primary abilities of Thurstone at the first stratum.

According to Carroll, the work of Spearman and Thurstone can be satisfactorily reconciled in a hierarchical model so that they only differ in the "relative importance they attributed to primary and second order factors" (1993, p.50). He noted that Thurstone originally developed a one stratum theory but that in later years Thurstone became interested in including group factors and a second order general factor:

> *Thurstone's assessment of the importance of group factors seems to be justified. (p.57)*

According to Carroll, many aspects of Vernon's hierarchical model were correct:

> *There is good evidence, for example, for clustering of variables around higher-order verbal-educational and spatial-mechanical factors, and for the domination of these factors by some sort of general factor. (p.60)*

Carroll described the Cattell-Horn model, as summarised by Horn (Horn 1986, Horn 1989), as

> *…a true hierarchical model covering all major domains of intellectual functioning. Numerous details remain to be filled in through further research, but among available models it appears to offer the most well-founded and reasonable approach to an acceptable theory of the structure of cognitive abilities. The major reservation I would make about it is that it appears not to provide for a third-order g factor to account for correlations among the broad second-order factors. (p.62)*

Carroll (1993) also claimed that there is a fairly close correspondence between his view of intelligence and Gardner's theory of multiple intelligences. Carroll identified 20 domains of cognitive ability in which he grouped factors:

> *That is, the first nine of these groups are regarded as true cognitive abilities in the sense of being relatively fixed, long-term attributes of individuals respecting the kinds of cognitive tasks they can and cannot perform, with varying degrees of success, at a particular stage of development. (p.137)*

Table 6 provides a general outline of the three-stratum structure of the major cognitive abilities that have been identified in Carroll's survey. According to Carroll, his analysis of abilities of several orders and strata offers insight into the structure of abilities and can be the basis for a scientific theory of cognitive abilities.

In Carroll's model a factor G at Stratum 3 (3G) dominates a series of broad abilities at Stratum 2. Carroll found consistent evidence of a general factor at Stratum 3 of his model:

> *There is abundant evidence for a factor of general intelligence, G (or 3G), found at the highest order (usually 2 or 3) of analysis for a given data set and thus at stratum III, that dominates factors or variables that emphasize the level of difficulty that can be mastered in performing induction, reasoning, visualization and language comprehension tasks. There is also some evidence that the G factor is likely to be correlated (though at a low level) with measures of speed of information processing and capacity of working memory.*

Carroll warned that inadequacies of available data mean it is not advisable to attempt to assign numbers or coefficients to indicate how strong this domination by Stratum 3 is, in the case of each second-stratum factor. According to Carroll, the strength of domination is roughly indicated in Table 6 by the closeness of each box representing a second-stratum factor to the left hand margin. This places fluid and then crystallised intelligence as closest to general intelligence. At a lower order of analysis, or at Stratum 2, a number of broad ability factors can be distinguished:

- language,
- memory and learning,
- visual perception,
- information-processing, and
- knowledge of certain general domains.

The stratum factors listed under each second-stratum factor are those regarded as being most likely to be dominated by the respective Stratum 2 factor.

Although Carroll developed a hierarchical model of cognitive abilities that includes a concept of *g*, he was agnostic about its impact or the significance of this general factor. Other theorist have continued to advocate the importance of a concept of *g* or IQ, but the notion of *g* and IQ has also drawn significant criticism.

Table 6. Carroll's Three Stratum Structure of Cognitive Abilities

Stratum 3							
General Intelligence 3G							
Stratum 2							
Fluid intelligence 2F	Crystallised intelligence 2C	General memory and learning 2Y	Broad visual perception 2V	Broad auditory perception 2U	Broad retrieval ability 2R	Broad cognitive speediness 2S	Processing speed 2T (RT decision speed)
Stratum 1							
Level factors General sequential reasoning (RG) Inductive (I) Quantitative reasoning (RQ) Piagetian reasoning (RP) **Speed & level factors** Speed of reasoning (RE)	**Level factors** Language development (LD) Verbal (printed) language comprehension (RC) Lexical knowledge (VL) Reading decoding (RD) Cloze ability (CZ) Spelling ability (SG) Phonetic coding (PC) Grammatical sensitivity (MY) Foreign language aptitude (LA) Communication ability (CM) Listening ability (LS) Foreign language proficiency ((KL) **Speed & level factors** Reading speed (RS) Oral production and fluency (OP) Writing ability (WA)	**Level factors** Memory span (MS) **Speed & level factors** Associative memory (MA) Free recall memory (M6) Meaningful memory (MM) Visual memory (MV) Learning ability (L1)	**Level factors** Visualisation (VZ) **Speed & level factors** Spatial relations (SR) Closure speed (CS) Flexibility of closure (CF) Serial perceptual integration (PI) Spatial scanning (SS) Perceptual speed (P) **Miscellaneous** Imagery ((M) Length estimation (LE) Perception of illusions (IL) Perceptual alternations (PN)	**Level factors** Hearing & speech threshold (UA, UT, UU) Speech sound discrimination (US) General sound discrimination (U3) Sound frequency discrimination (U5) Sound intensity duration discrimination (U6) Musical discrimination & judgement (U1, U9) Resistance to auditory stimulus distortion (UR) Temporal tracking (UK) Maintaining & judging rhythm (U8) Memory for sound patterns (UM) Absolute pitch (UP) Sound localisation (UL)	**Level factors** Originality/ creativity (FO) **Speed & level factors** Ideational fluency (F) Naming fluency (NA) Associational fluency (FA) Expressional fluency (FE) Word fluency (FW) Sensitivity to problems (SP) Figural fluency (FF) Figural flexibility (FX)	**Speed factors** Rate of test taking (R9) Number facility (N)	**Speed factors** Reaction time (R1) Choice reaction time (R2) Semantic processing speed (R4) Mental comparison speed (R7)

In reviewing Carroll's work, Spearitt (1996) saw significant implications in the model for teaching and learning, which included the:

- identification of learning difficulties;
- adaptive teaching; and
- training of Stratum 1 abilities.

Spearitt also saw in Carroll's model implications for research and curriculum development and the organisation of classrooms, schools and school systems.

The work of Carroll is significant and deserving of attention, but it remains rather opaquely technical, and it seems to have had had little or no impact or influence on educators, let alone the public in general. Too much of Carroll's analysis is of theoretical rather than practical interest, and he struggles to give it real world meaning and relevance. While they may account for differences in performance, Stratum 2 factors of broad memory, perception, retrieval and speed do not have much practical meaning, or if they do, Carroll does not communicate the significance of these factors.

Carroll's tireless analysis has led to the significant findings that there are three key abilities: verbal, abstract reasoning and spatial/visual abilities. Carroll also endorsed Cattell's notion of fluid and crystallised abilities, where abstract and spatial/visual abilities are the most fluid, and verbal ability is the most crystallised.

Factor analytic research has distinguished the three key factors of verbal, abstract/logical and spatial/visual. All factor analytic theories have a concept of verbal ability. For Carroll verbal ability is the residual after removal of the fluid reasoning ability. A non-verbal, abstract and logical reasoning ability is consistent across the psychometric models. A non-verbal spatial visualisation ability is also consistent across the models. The distinction between fluid and crystallised intelligence has been very influential.

The psychometric approach to defining cognitive abilities has been subject to considerable challenge and debate. The arguments advanced by supporters of *g* or IQ, like Jensen (1998) and Gottfredson (1997, 1998), have excited scepticism from Sternberg and active repudiation by Gardner and Ceci. Arguments about the nature of cognitive abilities readily extend into arguments about the influence of cognitive abilities on various positive and negative social outcomes. Defining and assessing cognitive abilities has historically been involved with political arguments about differences in race, gender and class. Table 7 presents a comparison of the nature and scope of the 10 theoretical approaches to understanding intelligence.

A different approach

The preceding discussion has reviewed the development of ability theories within the psychometric tradition. The thesis of the generalist Spearman, the antithesis from the specialist Thurstone, and the synthesis of Cattell and Carroll, have been reviewed.

For the first 30 years the ability testing of ACER was within the psychometric tradition, but the development of cross-curricular tests at ACER in the 1960s diverged from the kind of cognitive testing of the psychometric movement sketched above. While the CSSE used psychometric methods, it did not test psychometric constructs. Unlike the laboratory type tasks of the psychometric tradition of differential ability testing, the CSSE was:

- stimulus based;
- curriculum related;
- real world or meaningful rather than abstract;
- concerned with higher-order thinking rather than basic processes and skills; and
- based on unique items prompted by the stimulus material rather than formulaic and algorithmic items.

The psychometric tradition used formulaic, laboratory type items testing separate, basic skills, but the kind of unique stimulus-based testing developed for the CSSE, aimed at unique and complex items that required artful test writing. Here are a couple of examples that typify artful test writing and should be distinguished from the formulaic tradition.

As with everything in the CSSE, Exhibit 9 and Exhibit 10 involve reading comprehension, but Exhibit 9 is concerned with images and Exhibit 10 is concerned with understanding the conceptual relationships presented in text and diagrams.

The questions in Exhibit 9 relate shapes, overall design and purpose. Candidates are to attend to the nature of the shapes, the way space is represented to give an overall impression.

Table 7. A comparison of the nature and scope of the 10 theoretical approaches to understanding intelligence

	Knowledge based	Structural Piaget	Psycho-metric	Information processing	Genetic biological	Multiple intelligence	Modular	Contextualist	Triarchic	Bio-ecological
Nature of assumptions										
Existence of a substantial general ability *g*	No	Yes	Mixed	Mixed	Mixed	No	No	No	Yes	No
Existence of special abilities	Yes		Yes	Mixed	Yes	Yes	Yes	Yes	Yes	Yes
Biological bases		Yes	Mixed		Yes	Yes	Yes		Yes	Yes
Transdomainal processes	No	Yes	Mixed	Yes	Yes	No			Yes	No
Context sensitive		No		No	No	Yes	No		Yes	Yes
Process v product oriented	Both	Process	Product	Process	Product	Product	Process	Both	Both	Both
Role of motivation						Yes	No	Yes	Yes	Yes
Developmental	Yes	Yes	Mixed	Mixed	Mixed	Yes	No	No	No	Yes
Inductive/deductive balance	D	D	I	D	I	D	D	I	D	I
Scope of evidence										
Sociological					Mixed			Yes	Yes	Yes
Anthropological	Mixed				Mixed	Yes		Yes	Yes	Yes
Historical							Yes			Yes
Genetic			Mixed	Mixed	Yes	Yes	Yes	Mixed	Yes	Yes
Educational		Yes						Yes		Yes
Micro level processing	Mixed	Yes		Yes			Yes		Yes	Yes
Macro level processing	Yes	Yes	Yes	Mixed	Yes	Yes	Yes	Yes	Yes	Yes
Role of knowledge	Yes			Mixed				Yes	Yes	Yes
Reaction time			Yes	Yes					Yes	Yes
Positive manifold			Yes						Yes	Yes
Life course		Yes	Mixed		Mixed					Yes

It is worth noting that from the current view of test developers, Item 34 is weak because it is a negative question. Negative questions are not used in recent item writing because they add a spurious degree of difficulty in changing the task from finding the right answer to finding the wrong answer.

The strength of Exhibit 9 is the way that it asks candidates to recognise the impact of graphic design. The shape of the Barchan dune discussed and drawn in Exhibit 10 is genuinely difficult to understand. As well as being surprising in its distinctive form and ability to move, it is difficult to understand from both the text and the diagrams the way the wind direction shapes the different slopes. The matter is not gratuitously complex, it is essentially counter intuitive and confusing.

Both of these units show something of the range of CSSE Humanities.

Exhibit 9.
Travel posters, Humanities 1971
Items 31-38

The questions in this unit are based on two travel posters.

31. Below are several statements about lines. Which one of them most accurately typifies the lines used in Poster I?
A. A combination of straight and curved lines with neither kind predominating.
B. Jagged lines generally running vertically up and down the poster.
C. Curved lines forming a variety of shapes and patterns in the poster.
D. Straight lines forming a predominantly angular pattern.*

33. The outlines of the various features presented in Poster I are most accurately described as
A. becoming blurred when one feature overlaps another.
B. relatively indistinct for the features in the distance but more clear-cut for those in the foreground.
C. sharply defined for all the features presented in the poster.*
D. relatively distinct for the features in the distance but less clear-cut for those in the foreground.

34. Look at the way the various shapes in Poster II are grouped. Which one of the following is not part of what is conveyed by this grouping?
A. The height and size of the buildings.
B. The number of windows used in the buildings.*
C. A sense of perspective.
D. The density of the buildings in this area.

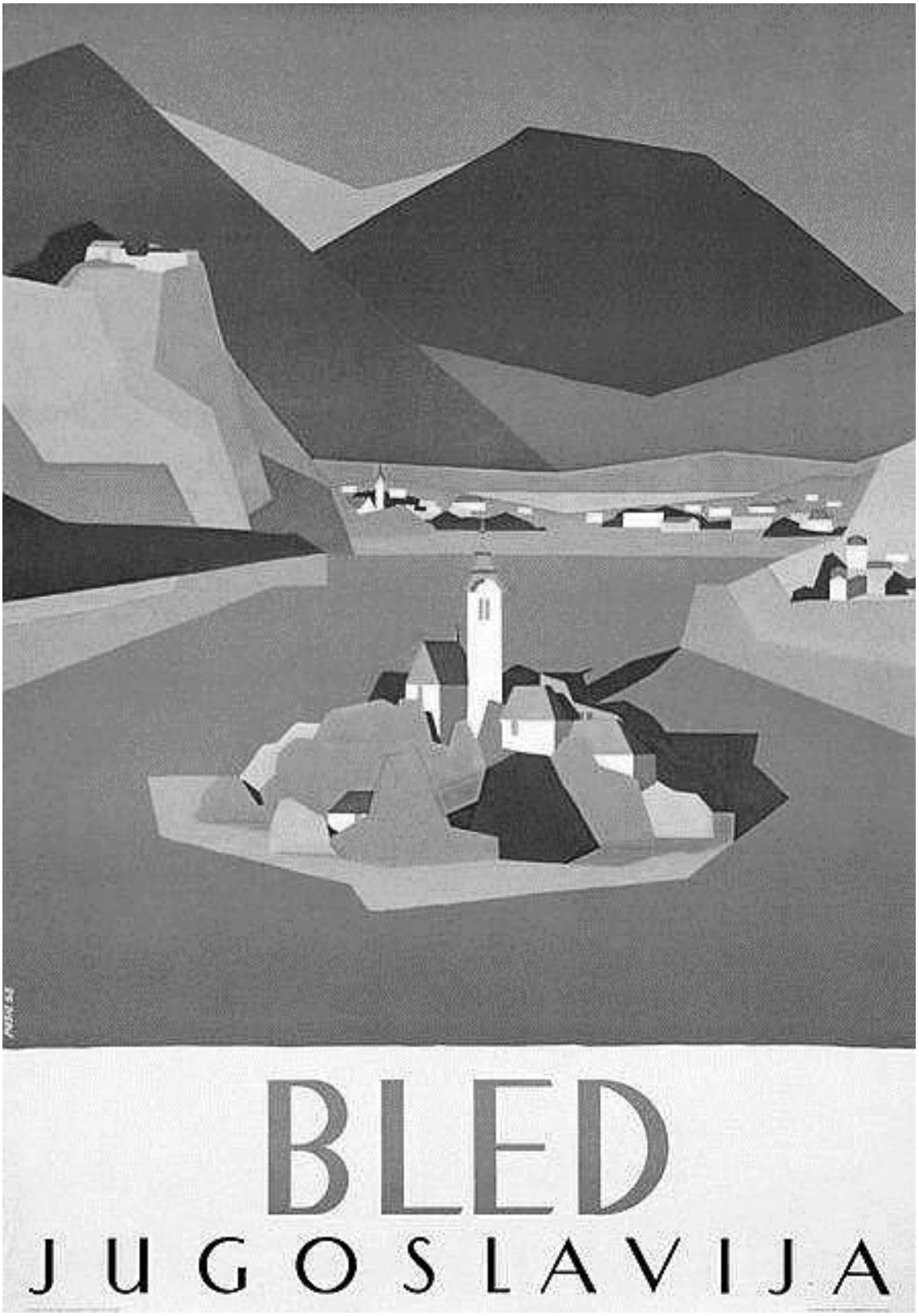
35. The contrast in the kind of holiday offered by each poster is primarily conveyed by their use of
A. simplified lines.
B. lettering.
C. geometric shapes.
D. colour.*

37. Which one of the following phrases best sums up the kind of vacation offered by Posters I and II respectively?

	Poster I	Poster II
A.	Local colour.	Time to explore New York.
B.	Crowds of people.	A variety of outdoor amusements.
C.	Tranquil relaxation.	A continuous round of entertainment.*
D.	A wide variety of different activities. Hustle and bustle.	

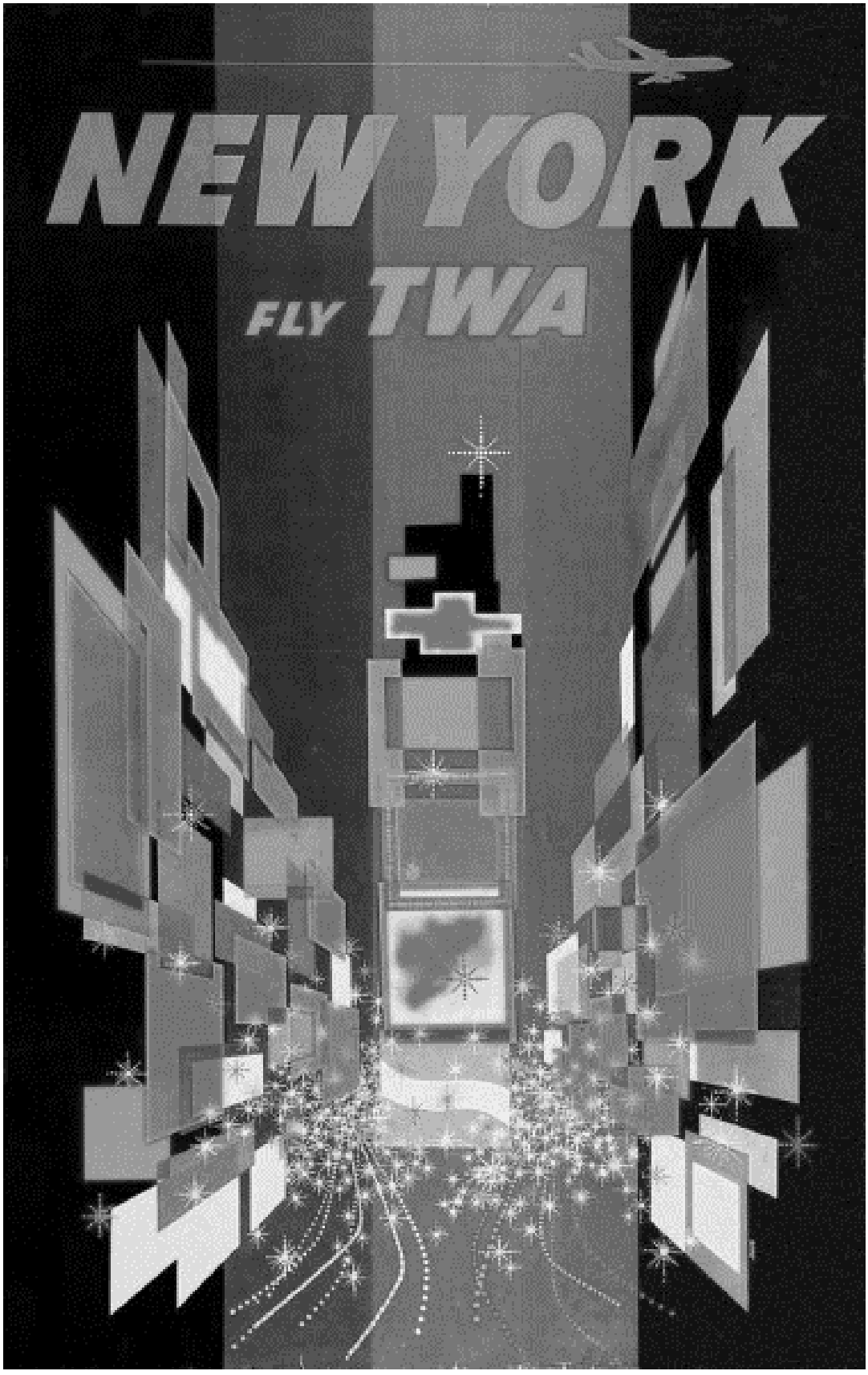
38. Below are four statements about posters. Which one of them is most characteristic of both Poster I and Poster II?
A. A poster is a message dramatically expressed in visual terms.*
B. A poster must convey several different ideas at one and the same time.
C. A poster conveys its point by means of elaborate and complex designs.
D. A poster must always represent realistically the place advertised.

Continued over page…



Poster I. Bled Jugoslavija - by Nesic (1958)

Continued over page…



Poster II. New York - fly TWA by David Klein (1956)

Exhibit 10.
The Barchan Dune, Humanities 1964
Items 81-87

The shape of the Barchan or Crescent Dune ... is distinctive. It may be described as a circular dome of sand, from the leeward side[1] of which a big bite has been taken, leaving a steep slope of loose flowing grains in the form of a hollow semi-circle. The Barchan advances down-wind as a crescent with its two horns, each tapering to a point of sand on the ground, reaching out in front of it. The horns of a single Barchan may be as far apart as 400 yards, its maximum height may be 70 to 100 feet, and its weight something up to half a million tons.

Individual dunes of this type can exist by themselves many miles from their nearest neighbours and in country otherwise quite free from scattered sand grains; but it is usual for them to grow in colonies reaching down-wind in a broad belt, separated by shorter and shorter distances from one another as one goes towards the source of the sand, that is, toward the windward end of the colony. Here the individuals are packed so close together that they run into one another, but in all cases the characteristic crescent bite is preserved.

The Barchan consists of two simple surfaces, the dome and the concave surface of loose collapsing sand round the bite. The dune keeps its simple geometrical shape intact with extraordinary persistence even while it is on the move, and while it is passing over such large obstacles as rocks, small hillocks, and villages.

Some innate tendency, some unknown effect of the interaction of wind and sand, keeps the ends of the two widely separated horns exactly level with each other. One can imagine some consciousness sitting on the domed top like a charioteer, checking each of them in turn in their blind advance, with a pair of invisible reins.

[1]leeward side is the side opposite to the direction from which the wind is blowing.

Bagnold, R. A. *Libyan Sands*. Hodder and Stoughton Ltd.

81. Using only the information contained in the passage, indicate which one of these descriptions best fits the area in which the Barchan Dunes are found.
A. A barren land, dominated by huge mounds of sand.*
B. A waterless waste, veneered by an uneven layer of sand.
C. A tropical wasteland, shunned by man and beast.
D. A desert wilderness, overwhelmed by sun and wind.

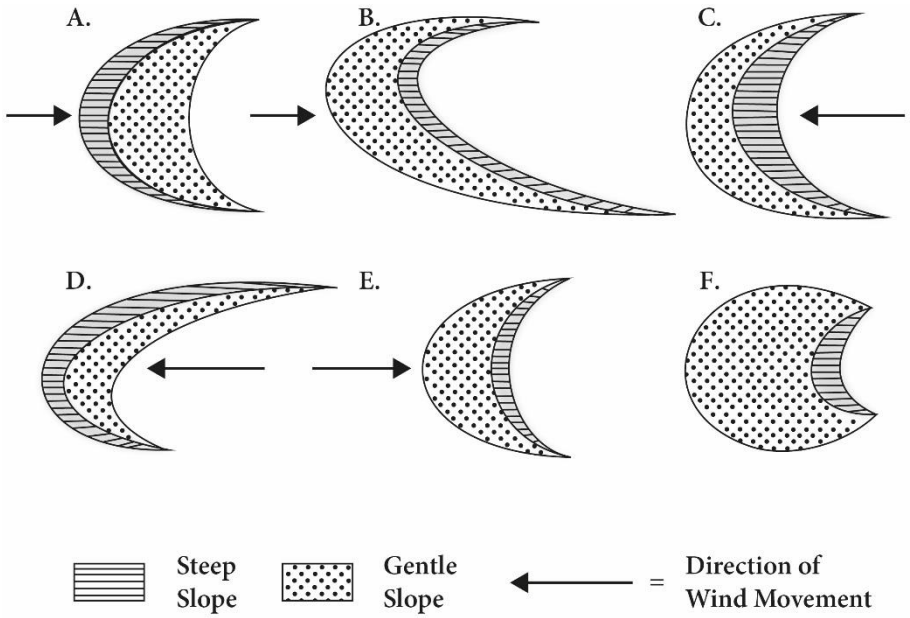
82. The chief impression the writer gives of the Barchan Dune is that it is
A. economically sterile.
B. an infertile mass of sand.
C. an almost living creature.
D. geometrically perfect.*

83. The term 'in their blind advance' means that
A. the ends of the Barchan are part of a heavy dust-storm.
B. no-one knows where the ends of the Barchan will finish up.
C. the ends of the Barchan, unlike human beings, have no eyes to guide them.*
D. there is no brain controlling the direction in which the ends of the Barchan will go.

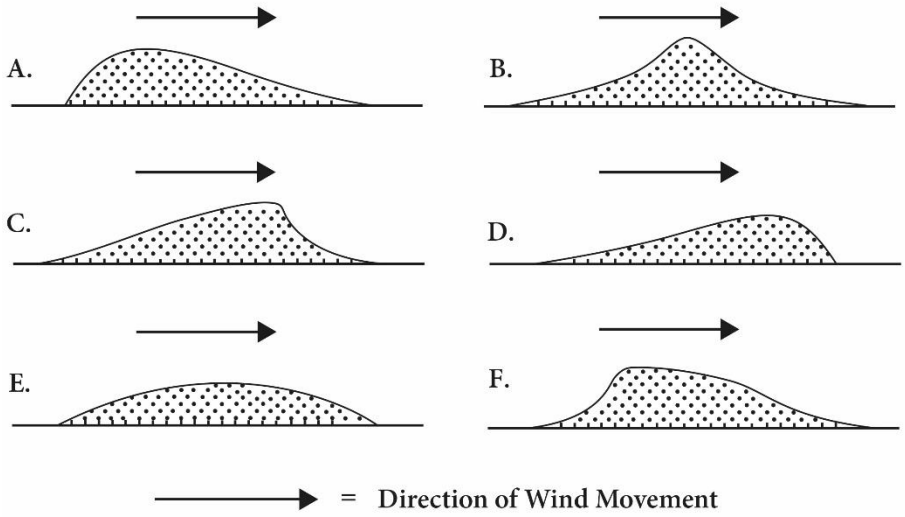
84. You are out in the desert, walking with the wind blowing on your back. You enter an area in which there are Crescentic Dunes. Within the next mile and a half you will expect to encounter
A. huge mounds of sand packed closely together but gradually decreasing in frequency.*
B. many large hillocks of sand clustered together.
C. a number of huge mounds of sand, gradually increasing in size.
D. the piling up of sand into evenly spaced domes from which large bites have been taken.

85. The ends of the dunes keep exactly level with each other. In the first sentence of the last paragraph, the discussion of this phenomenon by the writer is an example of
A. metaphorical explanation.*
B. superstitious belief.
C. intellectual honesty.
D. geographical speculation.
E. experimental method.

86. Which one of the following diagrams best represents a Barchan Dune as seen from above? E



87. If a straight cut were made through the middle of a Barchan Dune in the same direction as the prevailing wind, the section would look like which one of diagrams below? C



5

The CSSE, the SAT and the ASAT

This and Chapter 6 give some detail about the actual content and approach to multiple-choice test development in the **Commonwealth Secondary Scholarship Examination** (CSSE) and other cross-curricular tests produced at ACER. The subsequent discussion traces a dynamic tradition that grows and adapts, but that retains some distinctive characteristics initiated with the CSSE.

The terms of this discussion reflect the traditional test development culture of ACER. This culture sees test development as an art that entails subtle judgement about individual items and whole tests rather than treating test development as a production line of standardised items. The art of item writing is discussed in some detail in Chapters 6, 9 and 10. This chapter aims to review the cross-curricular multiple choice items developed by ACER for the CSSE and the **Australian Scholastic Aptitude Test** (ASAT) to the current day, and to compare them with the **Scholastic Aptitude Test** (SAT) of the **College Entrance Examination Board** (CEEB) in the United States.

The first CSSE test of 1964

The CSSE battery was eight hours of testing over two days. It was the longest, and in that sense it is the most comprehensive test ever developed by ACER. As detailed in Chapter 8, the current AST test is not as long, but in a sense it is broader than the CSSE. The structure of the first and the last administrations of CSSE are presented in Table 8.

At eight hours of testing, the CSSE is a very substantial and comprehensive battery. As suggested by the titles of the sub-tests, there is a similar emphasis on comprehension and interpretation in the science and humanities sub-tests. There were more humanities than science or quantitative items in 1964, and the number of items in the humanities and science test reduced over the decade as confidence in the effectiveness of the test grew. The writing test changed markedly over the 10 administrations, and these changes are examined in Chapter 7.

The CSSE aimed to test broad and generic rather than specific skills and abilities. It was stimulus based testing that involved meaningful thinking tasks. The CSSE was designed to produce an overall score. The battery was conceptualised as whole, while being related to and differentiated in terms of broad curriculum areas.

Table 8. First and last administrations of CSSE

CSSE 1964	Morning	Afternoon
Day 1	**Written expression** 2 hours 2 short composition (15 minutes) and 1 long composition (45 minutes)	**Comprehension and interpretation – Sciences** 2 hours 13 units and 74 questions
Day 2	**Quantitative Thinking** 2 hours 64 questions	**Comprehension and interpretation – Humanities** 2 hours 100 questions
CSSE 1973	**Morning**	**Afternoon**
Day 1	**Written expression** 2 hours 4 essays of 30 minutes	**Comprehension and interpretation – Sciences** 2 hours 60 multiple-choice questions
Day 2	**Quantitative Thinking** 2 hours 60 multiple-choice questions	**Comprehension and interpretation – Humanities** 2 hours 86 multiple-choice questions

The SAT and the ASAT

The **SAT Reasoning Test** (which was once called the Scholastic Aptitude Test) is the oldest continuous academic testing program in the United States. The SAT had been in existence for more than 30 years when the CSSE was developed. When the CSSE type testing was extended into tertiary selection that test was named the **Australian Scholastic Aptitude Test**, but a comparison between the SAT and the ASAT shows that the ASAT was in the CSSE rather than the SAT tradition.

The not for profit corporation based on independent colleges called the CEEB that owns the SAT R describes it as follows.

> *What is the SAT®?*
>
> *While high school grades are a very useful indicator of how students will perform in college, there is great variation in grading standards and course rigor within and across high schools. More than 80 years ago the College Board created the first standardized college entrance test to help colleges and universities identify students who could succeed at their institutions and to connect students with educational opportunities beyond high school.*
>
> *Today, the SAT is the benchmark standardized assessment of the critical reading, mathematical reasoning, and writing skills students have developed over time and that they need to be successful in college. Each year, more than two million students take the SAT. Nearly every college in America uses the test as a common and objective scale for evaluating a student's college readiness.* *http://professionals.collegeboard.com/testing/sat*

The SAT has had and still has a great influence on education in the United States because of the role it plays as one of a few common assessments available in a system of locally controlled schools.

In 2002 Lawrence, Rigol, Van Essen, and Jackson produced a College Board Report entitled *A Historical Perspective on the SAT 1926–2001*. The story told by Lawrence et al. shows that the SAT was fundamentally shaped by the growth in the United States of psychometric testing between the First and Second World Wars. While they used essay examinations in its earliest years, the new forms of psychometric testing soon came to influence American college entrance

testing. The curriculum related essay questions used in earlier versions of the SAT gave way in the 1926 version to nine subtests: seven with verbal content (definitions, classification, artificial language, antonyms, analogies, logical inference, and paragraph reading) and two with mathematical content (number series and arithmetical problems). In 1930, the SAT was split into two sections designed to assess 'verbal aptitude' and 'mathematical aptitude'.

While recognising that the SAT had its roots in the psychometric intelligence testing, Lawrence et al. see this approach as repudiated by 1946.

> *Critics of the SAT often point to its heritage in the intelligence tests of the early years of the last century and condemn the test on account of its pedigree, but it is worth noting that by 1946 those question types that were most firmly rooted in the traditions of intelligence testing had fallen by the wayside, replaced by questions that were more closely allied to English and language arts.*

Table 9 gives an overview of the changes in the composition of SAT after it moved away from the 'traditions of intelligence testing'. The nature of the SAT between the 1950s and the 1970s is of most interest for this discussion as it is the period in which the CSSE was conceptualised and developed into the ASAT.

Table 9. Overview of the composition of SAT Verbal

SAT Verbal	1958-74	%
Antonyms	18	20
Analogies	19	21
Sentence completions	18	20
Reading comprehension	35	39
Critical reading	7 passages	
Total verbal	90	
Total testing time	75 minutes	

SAT Verbal 1958 to 1978

According to Lawrence et al. (2002) the SAT of this period was inspired by language arts. In comparison with the test of 1926, 40 per cent of the test of 1958-74 was reading comprehension of the kind shown in Exhibit 11.

According to Lawrence et al. the first of these questions is an easy to medium-difficult example that "can be answered without making any complicated inferences", but it does "ask the test taker to make a connection between the text and his or her own life". The second question asks candidates "to evaluate and pass judgment on the passage". These are rather large claims for a pair of rather conventional reading comprehension items. The first question is asking what advice Johnson offers the reader, and the second is a logical matching exercise rather than an 'evaluation'. Compare these items with the first reading items in the first version of CSSE Humanities which are presented in Exhibit 4 and discussed on page 16.

Exhibit 11.
A reading comprehension example from the SAT, 1958-74

Talking with a young man about success and a career, Doctor Samuel Johnson advised the youth 'to know something about everything and everything about something'. The advice was good - in Doctor Johnson's day, when London was like an isolated village and it took a week to get the news from Paris, Rome, or Berlin. Today, if a man were to take all knowledge for his province and try to know something about everything, the allotment of time would give one minute to each subject, and soon the youth would flit from topic to topic as a butterfly from flower to flower and life would be as evanescent as the butterfly that lives for the present honey and moment. Today commercial, literary, or inventive success means concentration.

According to the passage, if we tried now to follow Doctor Johnson's advice, we would
(A) lead a more worthwhile life.
(B) have a slower-paced, more peaceful, and more productive life.
(C) fail in our attempts.
(D) hasten the progress of civilization.
(E) perceive a deeper reality. (Answer: C)

In which one of the following comparisons made by the author is the parallelism of the elements least satisfactory?
(A) Topics and flowers.
(B) The youth and the butterfly.
(C) London and an isolated village.
(D) Knowledge and province.
(E) Life and the butterfly. (Answer: E)

While 35 of the 90 items in the SAT verbal of 1964 were reading comprehension, more than 60 per cent of the items were the discrete psychometric item types of antonyms, analogies and sentence completions.

The antonym items tested vocabulary knowledge with the slight complication of having to be careful to match the opposite meaning.

VIRTUE:
(A) regret (B) hatred (C) penalty (D) denial (E) depravity (Answer: E)

These items were removed from the SAT in 1994.

Analogy items involved the matching of two pairs of words on the basis of a possible relationship of the words in a nominated pair. The matching is in the form of HAND is to PALM as FOOT is to ____? The formula was usually stripped back to use as few words as possible.

DOCTOR : HOSPITAL
(A) sports fan : stadium
(B) cow : farm
(C) professor : college
(D) criminal : jail
(E) food : grocery store (Answer: C)

The sentence completion items involved selecting appropriate words to fill gaps in a sentence. Such items are a kind of fill in the gap or cloze reading exercise. They involve vocabulary and grammatical knowledge with a kind of logical and grammatical inferencing that Lawrence et al. describe as having "an element of puzzle solving".

A …..… is a venerable leader ruling by …….. right.

mayor 1 patriarch 2 minister 3 general 4
paternal 1 military 2 ceremonial 3 electoral 4 (Answer: 2, 1)

The percentage of reading based items in the SAT verbal fell to 25 items in 1974. The reading items were increased to 50 per cent in 1994. According to Lawrence et al. the increase in reading items to 50 per cent of the verbal test was to 'send a signal to schools about the importance of reading'. One wonders what signal the vocabulary items, the verbal analogies and the sentence completion items had sent and continued to send to the schools of the United States.

Lawrence et al. describe the following "major changes to the verbal test" that were introduced in 1994.

- Emphasis on critical reading and reasoning skills
- Reading material that is accessible and engaging
- Passages ranging in length from 400 to 850 words
- Use of double passages with two points of view on the same subject
- Introductory and contextual information for the reading passages
- Reading questions that emphasise analytical and evaluative skills
- Passage-based questions testing vocabulary in context
- Discrete questions measuring verbal reasoning and vocabulary in context

The **Test of Standard Written English** was introduced into the SAT in 1974. It was a multiple-choice question test of language knowledge involving recognition of the conventions of standard written English, identifying unacceptable usage and choosing the best way of phrasing sentences. It was not until 30 years later that an optional 25 minute writing sample was added to the SAT in 2005. The current SAT writing test is 70 per cent multiple-choice question including error identification questions, sentence improvement questions, and paragraph improvement questions. The 25 minute writing sample is 30 per cent of the writing score.

The SAT of the 1950s to the 1990s was predominantly aimed at testing basic cognitive processes with discrete algorithmic items. The test is predominantly abstract and de-contextualised rather than contextualised or real world. There is an emphasis on word knowledge and mathematics that harkens after achievement rather than aptitude. The language testing is artificial and indirect.

In the 1990s the SAT evolved to give more emphasis to reading and less emphasis to vocabulary knowledge and mental gymnastics. The mathematics tests moved towards more authentic issues and problems. A direct test of writing was eventually included as 30 per cent of the Test of Standard Written English.

SAT Mathematics 1958 to 1978

Between 1942 and 1974 SAT Mathematics was made up of 70 to 80 per cent multiple choice. The other 20 to 30 per cent were called 'data sufficiency' items which were another kind of multiple choice (see Table 10).

Table 10. Overview of the composition of SAT Maths

SAT Maths	1959-1974	%
Multiple-choice question	42	70
Data sufficiency	18	30
Quantitative comparisons		
Student produced response		
Total testing time	75 minutes	

SAT Mathematics multiple-choice question

Lawrence et al. offer the following SAT Mathematics example which tests (1) knowledge of Pythagoras theorem in terms of the common 3, 4, 5 right angle triangle; (2) multiplying all terms with the common denominator to solve *b*; (3) deriving the answers by solving simultaneous equations. The two questions in Exhibit 12 are basic numerical knowledge and algorithmic technique.

Exhibit 12.
Pythagoras theorem, 1958-74

Example 1:

In Figure 1, if AC = 4, BC = 3, AB =

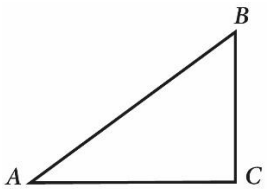


Figure 1

(Answer = 5)

Example 2:

If $4b + 2c = 4, 8b - 2c = 4, 6b - 3c = (?)$

(a) -2 (b) 2 (c) 3 (d) 6 (e) 10

(Answer = $b = 20$)

Data Sufficiency

Lawrence et al. offer the following example of a SAT data sufficiency item, presented in Exhibit 13. It tests basic knowledge of an isosceles triangle.

Exhibit 13.
Isosceles triangle, 1958-74

Directions: Each of the questions below is followed by two statements, labeled (1) and (2), in which certain data are given. In these questions you do not actually have to compute an answer, but rather you have to decide whether the data given in the statements are sufficient for answering the question. Using the data given in the statements plus your knowledge of mathematics and everyday facts (such as the number of days in July), you are to blacken the space on the answer sheet under

A. if statement (1) ALONE is sufficient but statement (2) alone is not sufficient to answer the question asked,

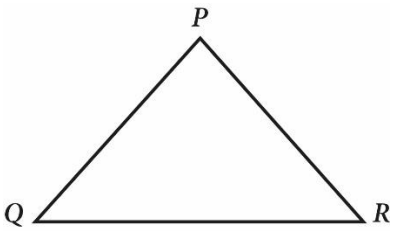
B. if statement (2) ALONE is sufficient but statement (1) alone is not sufficient to answer the question asked,

C. if BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient,

D. if EACH statement is sufficient by itself to answer the question asked,

E. if statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked and additional data specific to the problem are needed.

Example:



Can the size of angle P be determined?

(1) PQ = PR

(2) Angle Q = 40°

Quantitative comparisons

After 1974 the quantitative comparisons format was used in SAT Mathematics. Exhibit 14 presents two types of quantitative comparisons, offered by Lawrence et al. The first example tests

knowledge of the differences between mean, median and mode, while the second example is a puzzle.

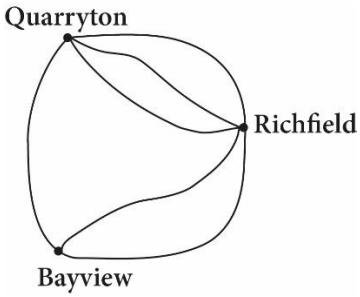
Exhibit 14.
Quantitative comparisons from the SAT, 1958-74

1. Seven students played a game and their scores from least to greatest are given below.
20, 30, 50, 70, 80, 90

Which of the following is true of the scores?
I. The average (arithmetic mean) is greater than 70.
II. The median is greater than 70.
III. The mode is greater than 70.

A. None
B. III only
C. I and II only
D. I, II and III
(Answer: B)

The figure below shows all roads between Quarryton, Richfield, and Bayview. Martina is travelling from Quarryton to Bayview *and back.* How many different ways could she make the round-trip, going through Richfield exactly once on a round-trip and not travelling any section of the road more than once on a round-trip?



A. 6
B. 10
C. 12
D. 16

Criticism of SAT

The SAT has been subject to on-going criticism for being inauthentic, abstract and de-contextualised. It is a brief test, and for much of its history it was a speeded test which candidates would have had rush to complete. It tends to test knowledge rather than reasoning, and procedural reasoning skills rather than conceptual thinking and higher-order skills.

There is an element of game playing and puzzling rather than authentic thinking in the SAT. It has been claimed that the knowledge-based and algorithmic nature of the SAT makes it teachable, and it is claimed that teaching the SAT has had a deleterious backwash on school curriculum in the United States. It has also been claimed that the lack of a substantial writing test until recent years has had an impact on the writing skills of students in the United States.

The contrast between the SAT and the CSSE/ASAT is marked. The SAT arises from the intelligence test tradition of discrete item testing, and it has only moved away from that tradition very slowly. As shown in Chapter 2, ACER was aware of the psychometric tradition of differential ability testing, and had adapted and produced tests of that kind, but it clearly eschewed that tradition in developing the CSSE.

It may seem unfair to compare the SAT as a fee-based tests of 150 minutes duration with a government funded battery of four tests of eight hours duration. The SAT is less than a third of the duration of the CSSE. The SAT would claim to have used discrete item testing so as to administer as many items as possible as quickly as possible. Such a view would seem to have had some impact on the earliest version of the ASAT, but ASAT soon became a three hour CSSE type test. When

it achieved the desired results, the discrete psychometric type items were removed from the early versions of the ASAT.

While the next chapter will contrast psychometric type test items with the kind of items used in the CSSE, Exhibits 15 to 19 are typical quantitative reasoning items from CSSE.

Exhibit 15.
The date of Easter, Quantitative ability 1968
Items 60 and 61

This year, Easter Day fell on March 26th, which is early; but it is not as early as possible, for unless world-wide agreement to the contrary occurs, Easter Day is and will be the first Sunday after the full moon which happens upon or next after the twenty first day of March. And if the full moon happens upon a Sunday, Easter Day is the Sunday after.

Now the average time from one full moon to the next is 29 days. 12 hours and about 44 minutes, but the time may vary by as much as about 6 hours.

60. When is the earliest that Easter Day can occur?

A. March 21.
B. March 22.
C. March 23.
D. March 24.
E. March 25.

61. What is the latest day in April on which Easter Day could fall?

A. 30th.
B. Anzac Day (25th).
C. 1st.
D. The Sunday following the second Saturday.
E. None of these.

Exhibit 16.
Election count, Quantitative ability 1966
Items 14-16

In a certain election there are five candidates P; Q, R., S, T of whom two are to be elected. Voters numbered their papers I to 5 in order of preference, and all voting papers were found to be satisfactory. The scrutineers in this election prepared the following table (leaving certain gaps):

		VOTES				
		1st	2nd	3rd	4th	5th
CANDIDATES	*P*	3	8	2	5	2
	Q	2		3		7
	R	7	5	3	3	2
	S	5	2		4	4
	T	3	1	7	4	5

A candidate is automatically elected if he receives 35 per cent or more of the first votes. If less than two candidates are automatically elected, then the other candidates are arranged in order of preference by awarding 5 points to a candidate for each first vote he obtains, 4 points for each second vote, 3 points for each third vote, and so on, and totalling the points so awarded.

14. The number of second votes received by candidate Q is

A. 2.
B. 4.
C. 5.
D. 8.
E. 20.

15. Which one of the following statements follows from the information given?

A. Only candidate P was automatically elected.
B. Only candidate S was automatically elected.
C. Only candidate R was automatically elected.
D. Both candidates R and S were automatically elected.
E. No candidate was automatically elected.

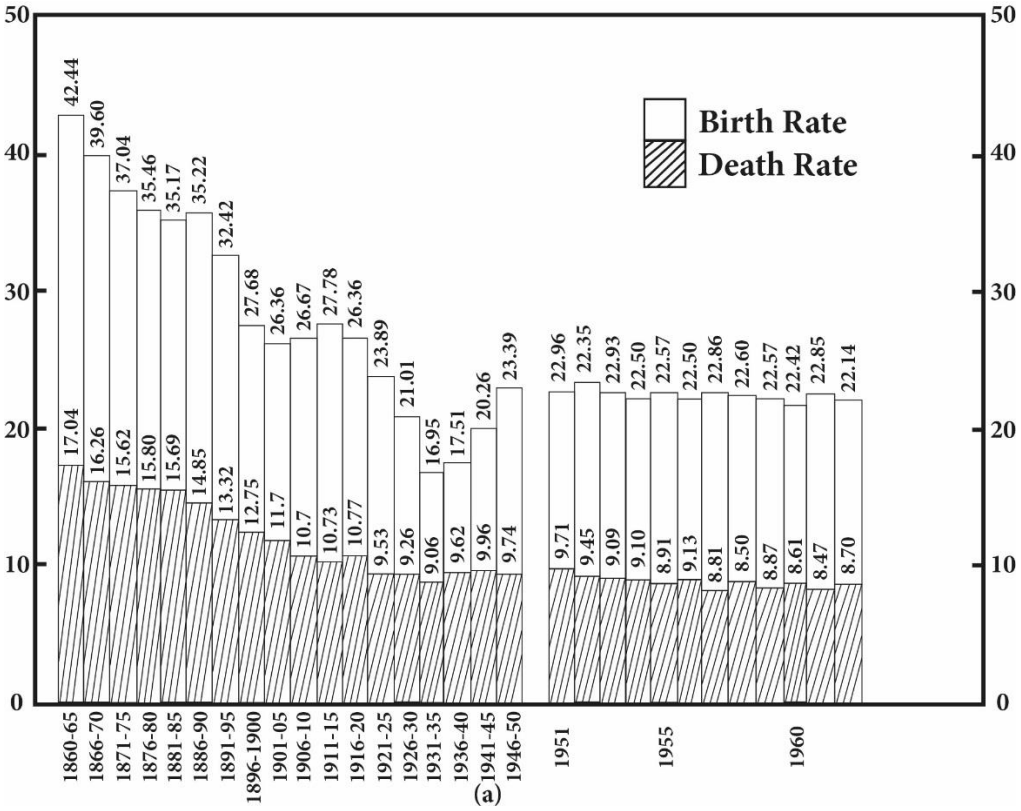
16. The two candidates finally elected were

A. Candidates P and R.
B. Candidates R and T.
C. Candidates R and S.
D. Candidates S and T.
E. Candidates P and S.

Exhibit 17.
Birth rates graph, Quantitative ability 1970
Items 4-6

CRUDE BIRTH AND DEATH RATES
AUSTRALIA 1861-

Numbers of Births and Deaths per Thousand of Mean Population
SOURCE: COMMONWEALTH YEAR BOOK



(a) EXCLUDES DEATHS OF DEFENCE PERSONNEL

4. In which of the following years was the crude birth rate the greatest?

A. 1951 C. 1954
B. 1952 D. 1955

5. In which of the following years was the natural increase in population the greatest?

A. 1954 C. 1960
B. 1959 D. 1961

6. In the period 1921 to 1950, the years 1931-35 had both the lowest death rate and the lowest birth rate. Of the following, which is the most plausible explanation of this low death rate?

A. There were fewer old people dying.

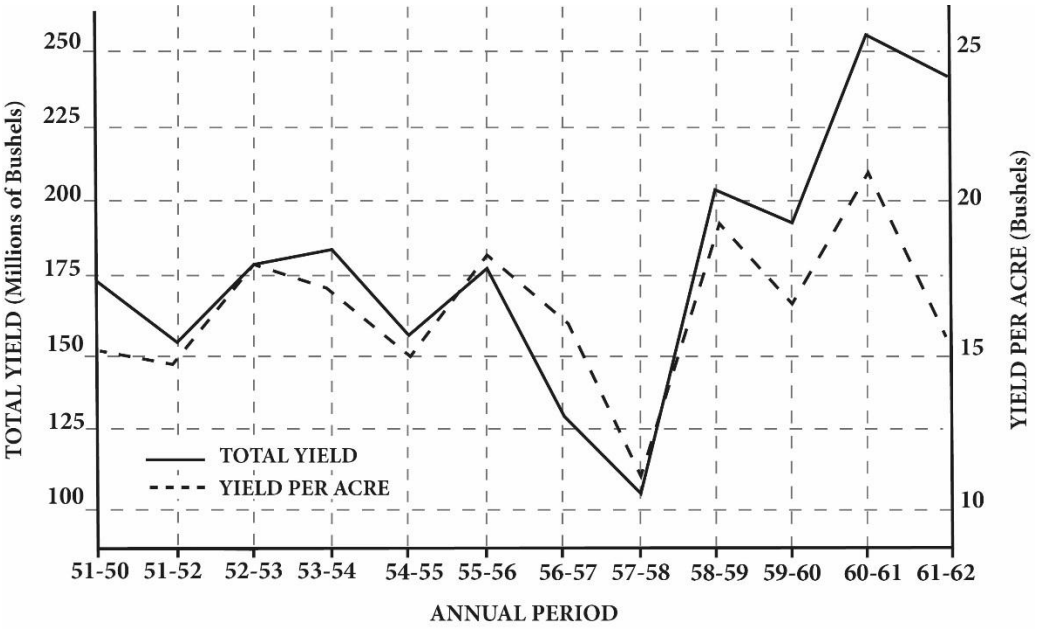
B. The total population was decreased.

C. The economic depression (of 1931-35) caused more sickness and death among middle-aged people.

D. Fewer babies were born and as a result fewer babies died.

Exhibit 18.
Wheat production graphs, Quantitative ability 1964
Items 40-42

Questions 40-42 refer to the following graphs based on the production of wheat in Australia for the period 1950-51 to 1961-62.



One graph shows the total yield of wheat for each annual period. The other graph shows the average yield of wheat per acre sown for each annual period.

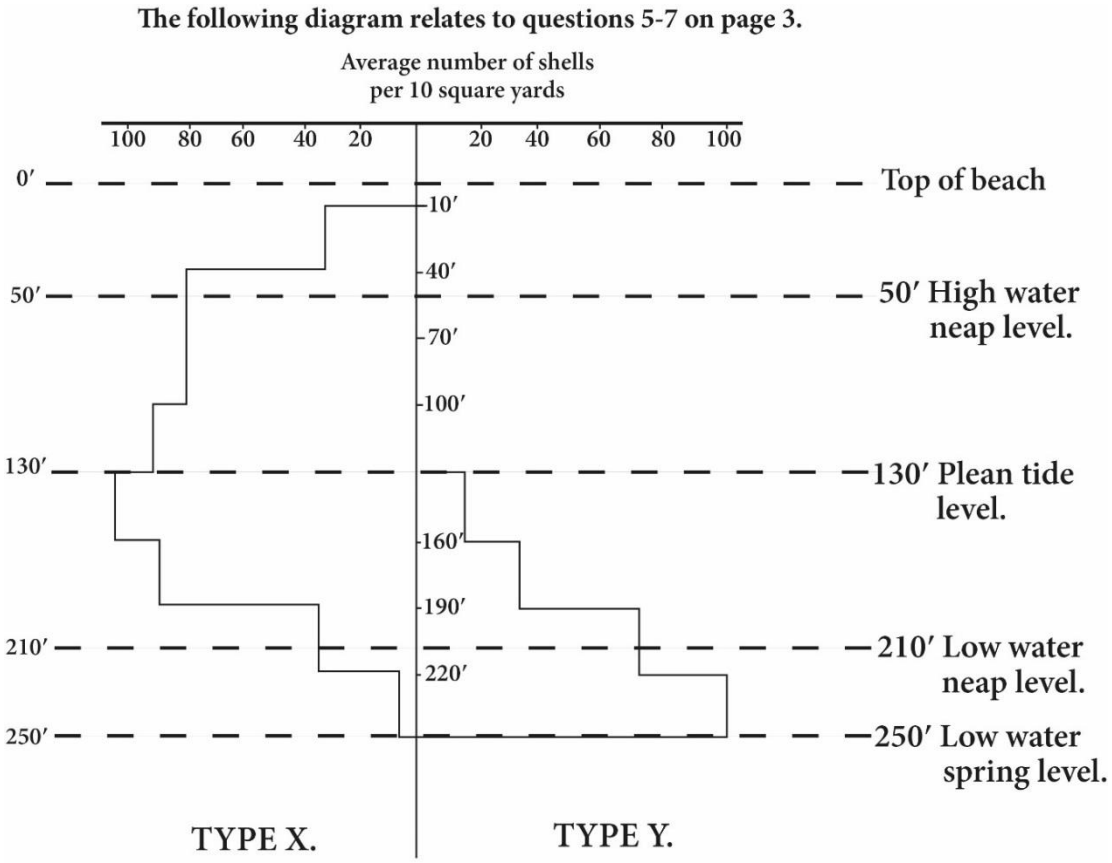
40. The average number of acres of wheat sown per year in Australia is approximately

A. 100,000 acres.
B. 500,000 acres.
C. 1,000,000 acres.
D. 10,000,000 acres.
E. 150,000,000 acres.

41. For which annual period was the number of acres of wheat sown the least?

42. For which annual period was the number of acres of wheat sown the greatest?

Exhibit 19.
Types of shell on a beach figure, Quantitative ability 1966



The diagram represents the results of an experiment designed to determine the distribution of two types of shells (X and Y) on a beach. The distribution of shells was only measured to low water spring level. The distribution shown above can be taken as representative of the distribution over the whole beach.
Use the information given above to answer the following questions.

5. Type X shells are

A. probably not found in quantity below low water spring level.
B. more plentiful at low water levels than at high water levels.
C. found on the beach in about the same numbers as type Y shells.
D. evenly distributed over the whole beach.

6. Type Y shells

A. are distributed over the whole beach.
B. are found higher up the beach than type X shells.
C. do not appear to be found above mean tide level.
D. do not occur below low water spring level.

7. There will be more type Y shells than type X shells on the beach

A. between mean tide and low water neap levels.
B. at mean tide level.
C. between low water and high water neap levels.
D. at low water neap level.

6

Art or algorithm? A comparison of psychometric and cross-curricular test items

Proponents of standardised psychometric testing claimed for much of the 20th century that multiple-choice reasoning tests were 'objective' in comparison with 'subjectively' scored essay examinations. It was also common to claim (or imply) that the abilities tested were innate and independent of social and educational background. In the 21st century few would make such claims. It has also been claimed by critics that psychometric style test items are coachable in a way that CSSE style tests are not.

The advocates or defenders of multiple-choice tests of reasoning now recognise that the general abilities tested are developed over time through social and educational experiences, although they would continue to claim they test general abilities rather than achievements in particular courses of study.

Multiple-choice ability tests have been criticised because they test 'lower order' skills with closed and reductive questions (Haladyna, 1997). According to this view multiple-choice question only test recall or logico-deductive and algorithmic kinds of thinking. The thinking tested is thought to be little more than a narrow kind of academic or test-taking intelligence. The most telling criticism of such multiple-choice tests is that their reductive narrowness has a deleterious backwash on teaching and learning (Neisser et al., 1996).

Although such criticisms have become more common and more vociferous in recent years, multiple-choice testing remains entrenched because it fulfils a necessary function in a cost-effective fashion.

Given that multiple-choice question can efficiently serve genuine purposes, it is unfortunate that such tests receive little substantive critique. There is little discussion of how such tests are better or worse, and how they might be improved. It is the intention of the following discussion to substantively examine some weak kinds of traditional psychometric multiple-choice questions, and to contrast them with some examples of the kinds of strong multiple-choice question that can be combined into broad and balanced generic skills tests such as the CSSE and ASAT.

Criteria for assessing multiple-choice items and tests

Strong multiple-choice items test authentic thinking and contribute to broad and balanced tests that are as meaningful and engaging to candidates as possible. Answering any multiple-choice question involves some kind of thinking, but the thinking becomes more authentic as it is closer to issues

and problems found in the real world. Such real-world thinking contrasts with the kind of test gymnastics that are typical of traditional psychometric testing of abilities (Sternberg, 1989; Ceci, 1990; Sternberg, 1996).

Authenticity has rarely been a consideration in the development of psychometric ability tests. With the intention of being 'culture-free', typical psychometric testing aims to have little relationship to the real world and to test 'abstract' intelligence (Cattell, 1987). There is little interest in breadth and balance in psychometric testing. Psychometric tests typically focus on one or a small number of things, and candidates are asked to do more or less the same thing 10, 20 or 30 times in a row. Whether there is any intrinsic interest or significance for candidates in the test items is not a consideration in traditional psychometric testing. Factorial homogeneity or statistical consistency of a particular construct is the primary aim of such testing, and multiple items of the same type increase such homogeneity.

Abstract intelligence

The 'abstract intelligence' posited by psychometrics is well represented in Raven's Progressive Matrices. The kind of items in Raven's matrices (as shown in Figure 5) are the archetypical abstract reasoning tasks. They are described as tests of 'fluid abilities' in contrast with 'crystallised abilities' that are developed through education and other experiences (Horn, 1994; Horn & Noll, 1997).

Items that test 'fluid intelligence' commonly involve abstract symbols and reasoning about visual patterns and relationships. It is argued by some that such items are the best test of intellectual ability in that they are not curriculum based and they are non-verbal. Such abstract, fluid reasoning contrasts markedly with thinking in language and through language about the world.

Abstract reasoning tests like Raven's matrices aim to test a schematic kind of formal and logical reasoning. Such items aim to minimise knowledge by being 'abstract': that is by having no relationship to the world. They do not test the kind of plausible reasoning (Hodgson, 1993) we usually use in thinking in and about the world. This abstract intelligence is the opposite of the traditional concept of social intelligence, Gardner's personal intelligences (Gardner, 1985) or the 'emotional intelligence' posited by Mayer and Salovey (1989; Mayer et al., 2000). Gardner's personal intelligences and 'emotional intelligence' have drawn attention in recent times because they contrast markedly with this 'abstract' intelligence.

Abstract intelligence tests present symbols that have no relation to the human world, and as a result such intelligence does not deal with matters that have any emotional charge or any social or personal meaning. Abstract intelligence is unemotional and impersonal processing of information (it is what a computer can do), and hence it contrasts with thinking about socially meaningful and emotionally charged issues about which an individual may have quite personal feelings and ideological commitments.

Issues that involve social, personal and emotional intelligence have meaning in the world and they cannot be represented as formal, logical reasoning. They involve different kinds of thinking, or a range of kinds of thinking. The notion that there are different kinds of thinking (rather than different kinds of information processing in different modalities) is not to be seen in much of the psychometric tradition. The greatest weakness of the psychometric view of thinking is that it has not managed to nuance the notion of reasoning into any significant differentiation other than abstract, spatial, mechanical and verbal.

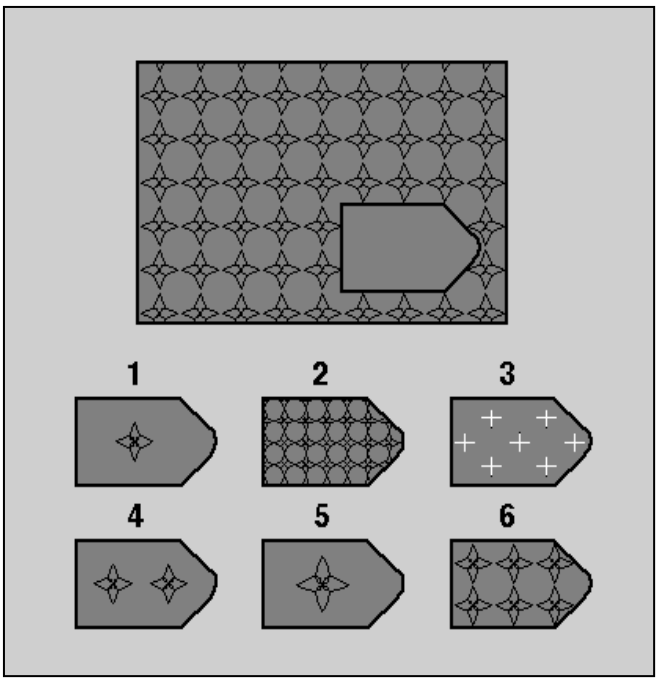


Figure 5. A typical item from Raven's Progressive Matrices

Verbal reasoning

Along with abstract reasoning, the other major component of the psychometric view of ability is the notion of verbal ability. Although the term 'verbal reasoning' is everywhere in cognitive psychology, it is rarely analysed and is usually treated as self-evident.

An exception is the compendious summation of psychometric views of ability presented by Carroll (1993). In his discussion Carroll offered 'verbal factors' as a kind of "baseline for the consideration of other factors":

> *It is often noted that the most frequently found factor in factor-analytic studies is a 'Verbal' (V) factor. (p.151)*

Although there have been significant attempts to differentiate verbal ability, Carroll concluded that knowledge of vocabulary and reading comprehension were the fundamental aspects of psychometric verbal ability. Carroll also noted that since this verbal ability was commonly a kind of general knowledge, and it was the crystallised intelligence against which fluid intelligence was distinguished. In essence, Carroll views verbal ability as a matter of how much language an individual knows, and how accurately an individual comprehends and interprets language.

Recall, reading and reasoning

Testing knowledge of word meanings is attractive because it is comparatively simple. Word knowledge can be directly tested, and there is a fairly clear hierarchy of word knowledge related to the frequency of word use (Carroll et al., 1971). Understanding words as concepts and in conceptual relationships with each other is more complex, and such understanding moves from recall of knowledge into reasoning with or about that knowledge.

Reading is a complex relationship of word knowledge, conceptual understanding and understanding of cultural context to make meaning. In its full complexity, reading is a kind of reasoning, but it is a largely tacit kind of reasoning, and the different aspects of reading may add up to different kinds of reasoning. Clearly the kind of reasoning involved in reading is significantly different from the abstract reasoning of fluid intelligence tests like Raven's Matrices or the Mill

Hill Vocabulary Scales, its accompanying test of 'verbal ability' through straight recall of vocabulary knowledge.

There is reason to question the assumption that vocabulary knowledge is an adequate proxy for reading ability, and there is reason to question the assumption that reading ability is an adequate proxy for plausible reasoning ability. Reading comprehension tests are commonly unclear about the extent to which they test reasoning, and they commonly conflate reading with reasoning.

It is arguable that reading is best seen as comprehension, and that reading literacy is best defined at levels of basic competence. It is unhelpful to confuse reading literacy with intelligence. One could be a poor reader who is very intelligent or a good reader (with wide word knowledge and language experience) who is not very intelligent. A reading test that is assumed or even designed to be an intelligence test will underrate the former and overrate the latter. It is the confusion of language knowledge and experience with intelligence that led psychometrics to notions of abstract and fluid intelligence. But it is not adequate to see abstract and fluid intelligence as the only way of limiting the part of language knowledge in intelligence. And it is not adequate to see reading comprehension (and is that what verbal ability adds up to?) as an adequate counterpoint to or foil for abstract intelligence.

The following test item might be called a test of reading comprehension, but it is clearly more than word knowledge and literal decoding. The language itself is quite simple, but the inferences required to answer the question are subtle. The following item asks candidates to determine what a statement means, and hence it is a reading item.

'I want to be what I was when I wanted to be what I am now.' *London Graffiti*

This statement expresses

A. longing for the past.*
B. acceptance of one's destiny.
C. excitement about the future.
D. satisfaction with the present.

The following item aims to evoke conceptual analysis and interpretation rather than reading.

'Liberty, Equality and Fraternity' is the motto of the French Republic.

Liberty is freedom from constraint.
Equality is having the same value or status.
Fraternity is the family relationship of brotherhood.

Which one of the following best describes the most likely relationship (X = likely to be in conflict, ♥ = unlikely to be in conflict) between the ideas of Liberty, Equality and Fraternity?

A. Liberty X Equality♥ Fraternity*
B. Liberty ♥ EqualityX Fraternity
C. Liberty ♥ Equality♥ Fraternity
D. Liberty X EqualityX Fraternity

Which one of the following best describes the most likely relationship (X = likely to be in conflict, ♥ = unlikely to be in conflict) between the ideas of Liberty, Equality and Fraternity?

A. EqualityX Liberty ♥ Fraternity
B. Equality♥ Liberty X Fraternity
C. Equality♥ Liberty ♥ Fraternity
D. EqualityX Liberty X Fraternity*

Answering these questions involves a complicated kind of conceptual analysis and reasoning. Candidates have to recognise that *Liberty* is potentially in conflict with *Equality* and *Fraternity*, while *Equality* and *Fraternity* are cognate or compatible. It is likely that one will have to give up some individual *Liberty* to promote general *Equality*. *Fraternity* is a kind of *Equality* which entails

a recognition of responsibility to others. This reasoning process to give answers to these items is not logically irrefragable, but it is plausible and convincing. No other relationship or interaction between *Liberty*, *Equality* and *Fraternity* is evident.

Kinds of verbal reasoning

Although the term verbal reasoning is commonly used in the psychometric tradition, it covers quite different kinds of tests. It has been common in ability tests for verbal reasoning to be (as in the SAT) a mixture of straight vocabulary knowledge, verbal analogies, grammatical and syntactical knowledge and reading comprehension. There is little or no reasoning in tests of vocabulary and grammatical knowledge, and verbal analogies and reading comprehension seem to be quite different kinds of reasoning.

The verbal analogies that are a standard part of psychometric tests are a kind of equivalent in language to non-verbal, fluid intelligence tasks. Verbal analogy items offer candidates a pair of words like *Doctor : Patient*, followed by *Lawyer : ?*. Candidates are asked to choose a word to pair with *'Lawyer'* that matches the relationship of doctor and patient.

What is striking about these exercises is that they are not the processes of analogising that are fundamental to thinking. They are not a process of finding a basis for a relationship between ideas or things that represent or illustrate characteristics. (Is the relationship between a doctor and a patient like that between a parent and a child?) Verbal analogy items are based on arbitrary comparisons, and one can see this arbitrariness by simply changing the pairs. For instance, the whole issue changes when the task becomes: *Doctor* is to *Patient* as *Taxi driver* is to *Passenger*. The match can still be made, but the grounds for the match has been changed. On the other hand, actual analogising would involve asking which is the better comparison for the relationship between a doctor and a patient: that between a lawyer and a client or that between a taxi driver and a passenger, or that between a parent and child, or that between the clergy and laity? The limited and arbitrary nature of verbal analogy items shows the weakness of the notions of verbal reasoning in the psychometric tradition, and the failure of that tradition to adequately differentiate kinds of reasoning.

In contrast with the so-called analogical reasoning items of the psychometric tradition, more authentic multiple-choice questions can be developed that actually test analogical reasoning by asking candidates to recognise meaningful rather than arbitrary relationships. The following item is a more authentic analogical reasoning task than the standard verbal analogy items.

> When we want to understand something or explain a view or feeling about something we often compare it, as an analogy, with something quite different that still has some kind of similarity.
>
> In the following questions you are offered an idea in **CAPITAL LETTERS** and then you are offered four things or ideas that it might be compared with. Of the four alternatives, choose the one that best compares with or illustrates the idea in **CAPITAL LETTERS**.
>
> Which one of the following would best suggest **CONFUSION**?
>
> A. a cave.
> B. a whirlpool.*
> C. a mountain peak.
> D. a river.

Quantitative ability

The notion of a general quantitative ability is commonly found in multiple-choice ability tests. Given the importance of mathematics as a body of knowledge it is not surprising to see this body of knowledge taken to be a general ability.

Both theoretically and practically, it is difficult to separate quantitative knowledge from quantitative ability. The factor analytic model of ability developed by Vernon (1993), for instance, shows quantitative ability as a sub-set of a more general verbal/educational ability. It is not uncommon for tests of quantitative ability (such as the SAT) to be little more than tests of mathematical knowledge.

Quantitative ability can be seen as a more general logico-deductive or formal reasoning ability in the manner of Howard Gardner's logico-mathematical intelligence. To be seen as logico-deductive and formal reasoning rather than mathematical knowledge, quantitative ability as a general ability should be carefully distinguished from knowledge of mathematics. It is usually not distinguished from knowledge of mathematics.

The unit about medal tallies (see Exhibit 37, p.112) is a clear example of quantitative reasoning rather than knowledge of mathematics. Some of the best CSSE Quantitative ability items ask candidates to think about numbers as systems in a new way. Several other examples are shown in the following three Exhibits.

Exhibit 20.
The measurement of population size, 1971
UNIT 3, Items 14-18

The measurement of population size is an important aspect of modern biological research. It is easier to estimate the number of people living in Melbourne than the number of cod in the River Murray. Over the past 40 years, several methods, four of which are described below, have been developed for estimating non-human populations. To illustrate the first three methods, imagine a large sealed room containing hundreds of flies. You wish to estimate the fly population of the room.

Method 1: Tracer technique
Obtain a known number of flies from the room, say 50, mark their backs with coloured paint, and release them again in the room. Then trap a sample on a piece of fly paper. If for example, 100 are trapped on fly paper and 10 of these have been painted, the original fly population is estimated to be 500.

Method 2: Ratio and proportion technique
Place an open cage of known volume (say 1 cubic foot) in the room. Measure the volume of the room (say 100 cubic feet). After some time some of the flies will be inside the cage. Count the number of flies (say 25). Assuming the flies are evenly distributed throughout the room the fly population is estimated to be 2,500.

Method 3: Modified tracer technique
Determine the proportion of male to female flies (say 2 to 1) by taking samples with fly paper, then release a known number of flies of one sex into the room (say 100 females). Now trap a sample of the new population on fly paper and compare the new ratio of male to female flies (say 1 to 1). The change in this ratio must have been due to the extra flies (in this example female flies) which were introduced into the room. The original population is estimated to be 300.

The fourth method below was originally developed to measure the number of baby fish in a fish-breeding pond.

Method 4: Fish-dipping technique
A small trap is lowered into the pond, and remains in place for a fixed time (say one hour). The trap allows fish to enter but not to leave. Suppose that 250 fish are trapped in this way. The trap is then removed, emptied into another pond, and then immersed in the original pond for a second one-hour sampling. Suppose that this time only 200 fish are trapped. Since the second sampling indicates a drop from the first sampling (here a 20 per cent drop), the population after the first sampling must have decreased in size. This decrease is due to the 250 fish removed by the first sampling. An estimate of the original population (1,250) can be made.

Two separate population size investigations will now be described,

Investigation Number 1
Salmon tend to swim together in large, well defined schools. In order to estimate the population of a migrating school of salmon, a number of fish were caught, marked by putting a small nick in their tail fins with a razor blade, and immediately returned to the population. When the fish arrived at their destination, a net was used to capture a random sample of them. The proportion of marked fish in the sample was then found, and used to estimate the population size.

Let N = the estimated number of fish in the migrating population at the beginning of the migration;
M = the number of fish originally captured, marked, and returned to the population;
R = the number of fish caught at random in the net at their destination; and
S = the number of fish which were marked, released, and recaptured.

14. Which one of the following can be used to calculate N?

A $N = \frac{SM}{R}$ **C** $N = \frac{RM}{S}$

B $N = \frac{R}{SM}$ **D** $N = \frac{S}{RM}$

15. Which one of the statements below, if true, would make the reasoning used in Investigation 1 incorrect?

A. Marking the tail fin of a salmon reduces its ability to swim to its destination.

B. A weak, unhealthy fish is just as likely to be caught for marking as a strong, healthy fish.

C. Marked salmon remain clearly distinguishable from unmarked salmon at the end of the process of migration.

D. The probability of capturing a marked salmon depends on the proportion of marked salmon in the population.

16. Subsequent investigation shows that about 25 per cent of the salmon which begin the migration do not arrive at their destination. Four students are discussing the effect this loss will have upon the estimate of N. Whose argument is the most sound in view of the information given?

A. Michele: 'As 25 per cent of the population has been lost, the calculated value of N underestimates the actual size of the population beginning their migration.'

B. David: 'The calculated value of N gives a good estimate of the size of the population arriving at the destination, but over-estimates by 25 per cent the number beginning their migration.'

C. Mario: 'Population losses will affect both the marked and the unmarked fish in the same ratio; the calculated value of N therefore gives a good estimate of the number of fish beginning their migration.'

D. Karl: 'Since population losses only begin to occur after the marked fish have been returned to the population, the calculated value of N still gives a good estimate of the number of fish arriving at their destination.'

Investigation Number 2

In a food-canning factory it was decided to estimate the fly population in order to establish a baseline for control measures. Every two hours, a fresh fly paper was hung in the centre of the factory. The first fly paper trapped and killed 32 flies; the second trapped and killed 24. During the investigation, all doors, windows, and other openings were kept tightly sealed.

17. This is an example of population estimation using the

A. tracer technique. C. ratio and proportion technique.
B. modified tracer technique. D. fish-dipping technique.

18. Which graph best shows the number of free flies in the factory over a period of several hours?

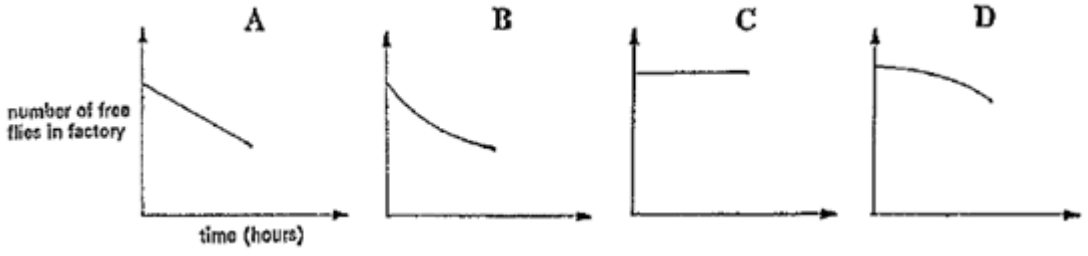


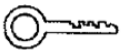
Exhibit 21.
Topographical transformations, 1971b
Items 53-54

Questions 53 and 54 refer to the following information:

Two surfaces are said to be topologically equivalent if each of them can be transformed into the other by stretching but without cutting or tearing. Assume that the solids referred to in Question 53 and Question 54 can be stretched or contracted in any direction without tearing.

53. Which one of the following solids has a surface that is not topologically equivalent to the surface of a drinking straw?

A flowerpot 

C key 

B match 

D chainlink 

54. Which of the following has surfaces that are topologically equivalent to the surfaces of three consecutive links in a chain?

A. Tyre, tube and bicycle wheel.

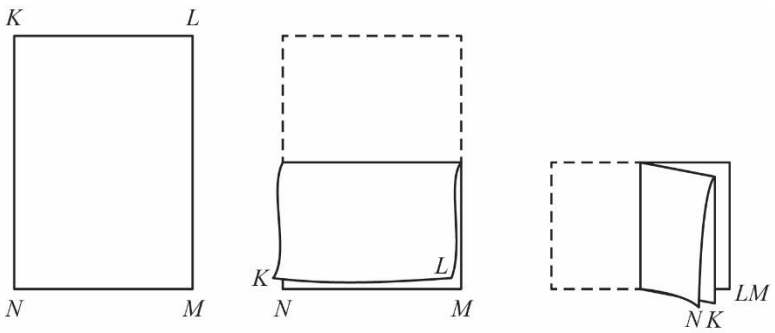
B. Bolt, washer and nut.

C. Two quoits and the peg on which they are thrown.

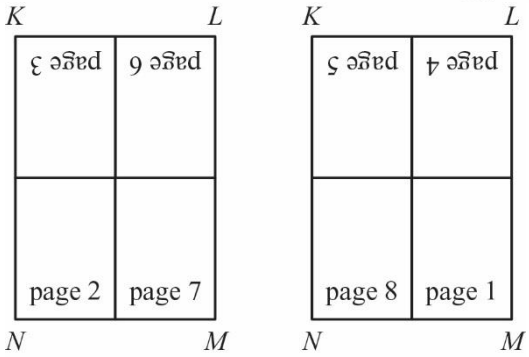
D. None of the above.

Exhibit 22.
Paper folding, Science 1971
Items 30-31

An eight-page newsletter consists of a single sheet printed on both sides and folded as shown:



The arrangement of a printed sheet is as shown:



30. Which one of the following statements is correct?

A When folded, some pages will be upside-down.

B When folded, page 5 will appear before page 4.

C The pages are arranged wrongly—page 2 should be on the back of page 1, page 3 on the back of page 2, and so on.

D There is no error in the layout.

31. Suppose that a sheet is printed as shown above and folded in the same way, but before it is folded, it is turned over so that the reverse side is uppermost instead of the front uppermost. In what order will the pages appear after folding?

A. 2 7 8 1 4 5 6 3 C. 7 8 5 6 3 4 1 2

B. 5 6 7 8 1 2 3 4 D. 8 7 6 5 4 3 2 1

Eschewing knowledge testing

In some sense all tests will involve knowledge, but this does not mean that there is no difference between a general test of formal reasoning ability and a test called quantitative ability that is actually (like the SAT) a test of mathematical knowledge. The status of knowledge in a generic skills test should be clear and carefully monitored.

A generic skills test should not aim to be a test of knowledge. In so far as it can, a generic skills test should not test what a candidate knows when s/he comes into the testing room. The propositional or declaratory knowledge needed to answer a generic skills question should be general knowledge, or it should be explained to the candidates. Difficult vocabulary or specialist knowledge should be glossed in the item stem or footnoted in a generic skills test.

Exhibit 23, about combustion and air, shows a kind of scientific thinking involving theory and hypothesising that is not a test of scientific knowledge. Candidates have to make logical inferences from the material and see how facts might be used as the basis for drawing conclusions.

Exhibit 23.
Burning in air, Science 1964

Two theories are advanced to explain the burning of material in air. Both theories assume that no substance can have negative mass or weight.

Theory X: During combustion the burning material unites with a certain component of the air. The capacity of air to provide this component is limited.

Theory Y: During combustion a substance escapes from the burning material into the surrounding air. The capacity of air to take up this escaping substance is limited.

Consider each fact below and decide how this fact, by itself, provides evidence for one theory or the other. Write

A. if the fact lends more direct support to theory X than to theory Y.
B. if the fact lends more direct support to theory Y than to theory X.
C. if the fact supports the two theories about equally well.
D. if the fact could not be used to support either theory.

N.B. - A statement is said to lend greater support to theory X than to theory Y if it can be more readily (i.e. with fewer or more reasonable additional assumptions) explained by theory X or if it contradicts Y but not X.

38. When a log of wood is burned the weight of ash that remains is much less than the weight of the original piece of wood.

39. An electric-light filament although completely enclosed in a small space will give off light and heat for long periods of time without any noticeable change.

40. A lighted candle in an enclosed jar of air stops burning before the candle is used up.

41. A gust of air will extinguish a glowing match.

42. A great deal of heat is given off during the combustion of coal.

43. The solid residue left when a certain metal burns in air has greater weight than the metal used up.

Similarly in Exhibit 24 about body temperature, candidates have to hypothesise and make conjectures rather than calculate. They are required to reason plausibly about cause and effect.

A strong generic skills test may well aim to introduce all candidates to issues they have little or no knowledge of, and ask them to come to terms with or make sense of these unfamiliar issues. In this sense a strong cognitive ability test will ask candidates to learn something from unfamiliar stimulus material, so that it is in a sense a test of the ability to learn.

Exhibit 24.
High temperature, Science 1967

Scientists make observations and then try to explain the cause of them. Here is one such observation:

Illness is often accompanied by a high body temperature.

Instructions - From the ten statements below choose the three statements you would investigate as being the most likely to explain the cause of this observed fact.

Write the number corresponding to your selections beside Questions 1 to 3 on the answer sheet. (The order does not matter).

Statements:

1. In all animals, many illnesses are characterized by high body temperatures.

2. It is obvious that if you have an illness you must have a high temperature.

3. The high temperature is nature's way of combating the illness.

4. Just as a car engine becomes overheated when it is overworked, so the human body does the same.

5. In some illnesses less blood flows in the surface of the body, so there is less heat loss.

6. This is very similar to the process of becoming hot with violent exercise.

7. Certain products of injured cells are carried to the brain, where they disturb the function of the heat-control processes.

8. The high body temperature makes the blood run faster and produces the fever.

9. The body contains a regulating device that increases the blood flow in the body under certain conditions of body malfunction, and this increases body temperature.

10. Many organisms causing illness flourish if temperatures are higher than normal body temperature.

Pushing the limits of the multiple-choice question format

A major weakness of the multiple-choice format is that it can only deal with issues for which there is a correct answer. In reality most thinking is not about definite problems that have one or even and number of solutions. Most thinking is a matter of plausible reasoning in which a thinker makes a decision based on incomplete or imperfect knowledge. A strong multiple-choice test will make some attempt to deal with the kinds of issues that are most difficult to test in the closed multiple-choice question format.

Sternberg has contrasted the artificial, structured and closed nature of academic/analytic problems with the real-life, unstructured and open nature of practical problems in the terms presented in Table 11 (Wagner & Sternberg, 1986).

In so far as it can, a strong ability test will push the limits of the multiple-choice question format. It will aim to test practical or real world problems and issues that involve plausible as well as logical reasoning. In contrast with multiple-choice question that have necessary or logical answers, plausible reasoning items will have the strongest answers the item writers can see to an 'ill-defined'

problem or an arguable issue. Such answers will be strong interpretations (that will never be beyond dispute), and they will be placed against distracters that can be eliminated on logical grounds, or on the basis of inconsistency with some stimulus material. Such plausible reasoning items will be inferential and interpretive rather than literal deductions. As we have seen, some of the best CSSE/ASAT items (such as Exhibit 1) are of this kind.

Table 11. Wagner and Sternberg's view of practical problem solving

Academic problems	Practical problems
Well defined	Ill-defined
Formulated by others	Unformulated
Necessary information provided	Additional information required
One correct answer	Multiple 'correct' answers
One method to obtain answer	Multiple methods to obtain answers
Dis-embedded from everyday experience	Embedded in everyday experience

Adapted from Neisser (1976) and Wagner and Sternberg (1986)

The following item in Exhibit 25, involves subtle inferencing about a mysterious, even ambiguous image by the great Saul Steinberg. The key to this item is not unarguably correct. It is not irrefragable. There may be a better key, but the item writers couldn't see it. The key is a strong and plausible inference. Such an item is escaping to a significant extent the limitations of the closed multiple-choice question format.

Exhibit 25.
Steinberg's thumb print

The image suggests that we

A. cannot escape conformity.

B. can effectively disguise ourselves.

C. appear to be anonymous but are unique.*

D. appear to be unique but are anonymous.

Escaping the limits of reading

Although for the time being multiple-choice tests will involve reading, strong cross-curricular tests should be explicit about the place of reading in the test.[3] While a test delivered in writing will inescapably test reading ability, the important issue is the extent to which the test can do more than

[3] According to the assumptions made here, an ideal multiple-choice test would be delivered simultaneously in a written and oral form. Such a mode of delivery will become possible in the future.

test reading as the retrieval of information. Some generic skills items should attempt to be conceptual reasoning rather than reading *per se*.

Shown in Exhibit 26 is one of the most famous and admired pieces of brushwork by the Zen master Sengai Gibon. It is generally called the universe, and is commonly taken to represent some fundamental Buddhist ideas.

The best way of testing more than reading is to keep the reading demands as low as possible for much of the test. There will be some high verbal material in a broad and balanced multiple-choice test, but much of the test will aim to be as low verbal as possible. A strong multiple-choice test would involve a substantial amount of reading and interpretation of diagrams and other visual material, such as that presented in the two dimensional chart of political views in Exhibit 27.

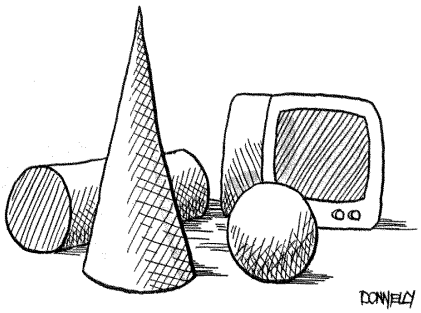
Rather than being high verbal, a strong multiple-choice test would aim to be conceptual. Although the test may be in language and may even be about language, a good multiple-choice test would be conceptual in that it would define words but may ask for inferences about the relationships between words, as in the Definitions of Meritocracy, Democracy, Aristocracy and Oligarchy unit (See Exhibit 28).

Exhibit 26.
Brushwork by the Zen master Sengai Gibon

Which of the following Buddhist ideas would make the most convincing interpretation of this image? The universe is:

A. an illusion.
B. eternally expanding.
C. empty and formless.
D. a unified and continuous whole.*

The following drawing by the New Yorker cartoonist Liza Donnelly seems a kind of response to the Sengai image.



The drawing above suggests that

A. geometric shapes are behind everything, even television.
B. television contrasts with other fundamentals.
C. geometric shapes are infinitely flexible.
D. television has become fundamental.*

Exhibit 27.
The two dimensional chart of political views

The chart below is a representation two political dimensions. A position on the chart for an individual (or a group) can be determined by a series of questions about Personal Freedom and Economic Freedom. Responses to the questions are scored (out of 100) to give a position on the chart.

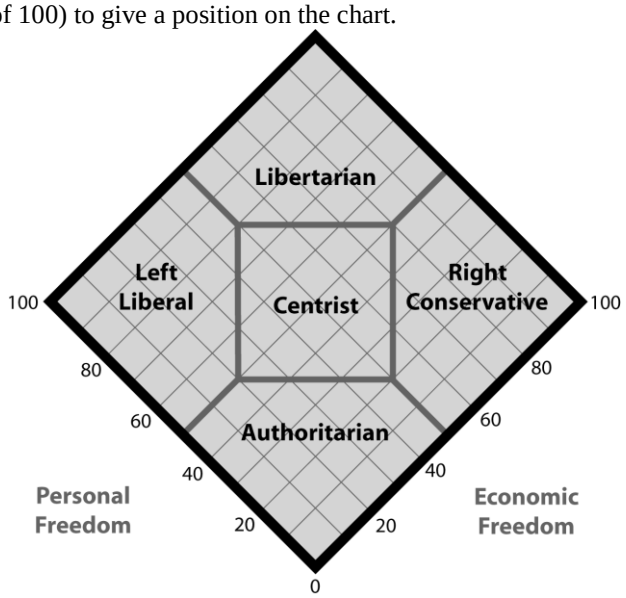


Figure 2. The Nolan Chart

The Chart suggests that Conservatives are

A. low on personal and economic freedom.
B. high on personal and economic freedom.
C. high on personal freedom and low on economic freedom.
D. low on personal freedom and high on economic freedom.

Which one of the following pairs of scores is closest to libertarianism?

	Personal Freedom	Economic Freedom
A	90	40
B	20	90
C	50	50
D	30	80

The Chart implies that

A. the right is more libertarian.
B. the left is more authoritarian.
C. both the left and the right can be libertarian.
D. both the left and the right are centrist.

The dimensions of the Chart vary from

A. more to less of one thing.
B. less to more of one thing.
C. less of one thing to less of another.
D. more of one thing to more of another.

Exhibit 28.
Definitions of Meritocracy, Democracy, Aristocracy and Oligarchy

Use the following definitions to answer the questions below.

Democracy is rule by all the people.	**Oligarchy** is rule by a small, elite group.
Meritocracy is rule by the most talented and capable	**Aristocracy** is rule by those born into a privileged class.

The definitions suggest that

A. aristocracy is more democratic than meritocracy.
B. meritocracy is more democratic than aristocracy.*
C. aristocracy can be democratic but oligarchy cannot.
D. oligarchy can be democratic but aristocracy cannot.

Which of the following forms of government are closest to each other?

A. Oligarchy and aristocracy.*
B. Meritocracy and oligarchy.
C. Aristocracy and democracy.
D. Meritocracy and aristocracy.

Meritocracy would be based on

A. who you are.
B. who you know.
C. where you come from.
D. what you know and can do.*

A foolish aristocrat is not supposed to succeed in a meritocracy because

A. success should be on merit.*
B. aristocrats must be meritorious.
C. the foolish should not be aristocrats.
D. the foolish should be treated as equal.

Meritocracy is

A. inconsistent with oligarchy.
B. inconsistent with democracy.
C. inconsistent with aristocracy. *
D. not inconsistent with all of the above.

The art of test construction

The psychometric tradition is focussed on standardised and algorithmic items that could ideally be decomposed into basic component processes, as in the componential analysis of intelligence described in Chapter 4. Cross-curricular test items are neither standardised nor algorithmic. They aim to be more or less unique responses to stimulus material. Similarly a broad and balanced cross-curricular test has a unique structure and balance. The construction of cross-curricular test is an art rather than an algorithm.

The contrast between test construction as art and algorithm is clear in the comparison of the diversity of the CSSE/ASAT items with the antonyms, analogies and sentence completions items of the SAT. An SAT form has a specified number of items of certain kinds. There is a technical choice about the statistical characteristics of the items, and there is some choice about which particular items are to be included, but otherwise the construction of a SAT form follows a specification and involves little judgement.

On the other hand, constructing a cross-curricular test like the CSSE or ASAT involves some careful judgements about breadth and balance. The kind of tables for reviewing a cross-curricular test shown at the end of Chapter 3 (see Exhibit 8) are of limited usefulness in this task. Such tables do not actually define or describe the texture and tenor of a particular cross-curricular test.

Clearly, the unspecified nature of a cross-curricular test can be problematic. The CSSE was an omnibus battery that could not be divided into specific kinds of items, as could be done with the SAT and other psychometric tests.

The CSSE/ASAT approach to test construction was challenging to some, and not without its critics. In the first research report on the CSSE, Thomson and Keeves noted that several 'obvious changes' had occurred in the CSSE since the objectives were published in 1967. According to Thomson and Keeves (1974), these changes included the following:

- a greater emphasis upon Arts topics in CSSE Humanities;
- eradication of the objective section in CSSE Written expression;
- a variety of non-verbal stimulus material in CSSE Written expression;
- the use of more mathematical content in the CSSE Mathematics paper which was originally designed as a test of Quantitative Thinking,
- extra disciplinary topics in CSSE Science; and
- a lesser emphasis on Social Science areas in CSSE Humanities.

Thomson and Keeves (1974) commented that:

> *Some may argue that these changes range from most desirable to being of dubious worth. Regardless of the validity of this argument, there is no doubt that changes have occurred without research being undertaken to investigate the effects of these changes.*

The lack of tightly structured specifications

Thomson and Keeves noted that there had been a good deal of attention given to the predictive validity of CSSE, but that other aspects of the validity of the test had not been examined. Thomson and Keeves examined the construct and concurrent validity of the CSSE by comparing performance on the test with performance on external examination results, school assessments, certain psychological tests of mental abilities, and a test of reading comprehension. They also collected background data about the age and sex of candidates, father's occupation and education and mother's education.

Thomson and Keeves had students take the CSSE and the following 'marker tests' to compare with performance on the CSSE.

- ACER Advanced Test AL-AQ – language and quantitative intelligence-type tests
- Ravens Standard Progressive Matrices – abstract reasoning
- Revised Minnesota Paper Form Board – spatial visualisation
- ACER Co-operative Reading Test – paragraph reading

The object of the comparison was to link an omnibus CSSE with specific constructs so as to explore and determine "the internal structure and the attributes purported to be measured by the CSSE".

Thomson and Keeves wanted to use confirmatory factor analysis to explore the internal structure of the CSSE, and to do so they needed a scheme for classifying the test items. Thomson and Keeves obtained descriptions of the procedures used in the construction of the CSSE papers from the persons chiefly responsible for them in 1972, but they found these descriptions "did little to explain

whether the objectives specified by Whitford were still being examined by these papers" (Thomson, 1974).

> *The constructors further stated that tight specifications for the CSSE papers were not used, and were not only impossible to state but undesirable for the testing of scholastic aptitude in the areas under consideration. To them the construction of the CSSE papers more paralleled an art than a science and should be viewed from that perspective. Tight specifications under these circumstances would have been too restrictive and thus were not used in the construction of the papers. The very general criteria used by the test constructors in compiling the 1972 tests are set out in Appendix L. These were not, and were not meant to be, amenable to empirical investigation.*

Nevertheless, Thomson and Keeves looked to the types and classification of items offered by Whitford in 1967 as the basis of a confirmatory factor analysis, but found it "was not only extremely difficult to use but proved to be of little help in interpretation". They noted the "lack of tightly structured specifications" for the CSSE, and they reluctantly used exploratory factor analysis on their data, but they found that the resulting clusters of items were very difficult to interpret. They drew the following conclusion.

> *With regard to the specifications of the papers, it would seem to the authors of this report to be most desirable that specifications should be set down and adhered to for the protection of both the test constructors and the examinees.*

This sentence had the following footnote: "We note that this view is not held by the test constructors".

The report records a clear difference about how the CSSE should be conceptualised and produced. Thomson and Keeves wanted a definite and tight test specification including more or less specific item types that could be demonstrated to have empirical meaning, and the specification would be used to develop items and to audit different test forms. This is the approach of the psychometric tradition to test construction. As mentioned in the footnote, a quite different view about how the CSSE test should be developed was presented in Appendix L of the report written by the test constructors.

In Appendix L the leader of the humanities test panel, Audrey Grant, offered three major criteria for constructing a CSSE Humanities test. The stimulus material was to:

- sample a broad range of subjects and media;
- be of interest and relevance to students; and
- have intrinsic quality and 'educational value'.

Grant (Thomson, 1974) described the Humanities items in the following terms.

> *The kinds of inquiry required of students in doing the Humanities paper arise from the choice and structuring of material in units and from the specific items. In relation to the given criteria - diversity, interest, quality, and overall balance of unit material - it cannot be overemphasized that those involved in writing the units and in the panelling process do contribute in a distinctly individual way. There is no predetermined content-process grid either for material or for question types. Such a grid would be seen as inappropriate and constraining in the English-humanities area. Recognizing the uniqueness or distinctiveness of each piece of written or pictorial material entails recognizing the appropriateness or otherwise of specific kinds of questioning about it.*

Grant defended the validity of the test on the basis of the development process.

> *In the absence of a content-process grid, or a clear cut specification of behavioural objectives, there is still a firm recognition of certain areas of questioning that are essential, and certain balances in emphasis and direction that need to exist.*
>
> *To stress the intuitive, individual, and qualitative factors that shape the tests provokes the demand for stating the grounds of validity. In the preparation of the Humanities paper the reply comes in terms of the panelling process combined with trial test data. The members of a panel bring to the test construction process a range of viewpoints and judgments which test constructors alone could not, and it is this panelling stage that enables a searching appraisal of the unit material and items, ensuring the discarding of unpromising leads, and the removal of ambiguities, unwarranted assumptions and value judgments resulting from the test constructors' approach.*
>
> *The panel provides a kind of validity in terms of expert opinion, both prior to and following the trial testing.*

Grant concludes the discussion with the following overview.

> *In sum, every effort is made to obtain as active an interaction between test writers, subject specialists and other educators as possible. This emphasis on interaction by educators of widely different interests has stemmed from the belief that the educational relevance and quality of the material used in the tests is as important as the statistical validity and reliability of the tests.*

In *From Testing to Teaching in the Humanities: Ideas and Resources for Use at the Upper Secondary School Level drawn from the Commonwealth Secondary Scholarship Examination 1971*, Grant and Withers (1973) presented a further description of the importance of the 'panelling process' (also see Appendix 8).

> *The preparation of test papers such as those used in the CSSE is not suited to individual examiners working in isolation. All resource material for the 1971 Humanities paper and all suggested questions and alternative answers were subjected to prolonged panel discussions. Each panel consisted of ACER staff members augmented by outside specialists, chosen not for their technical expertise in educational testing, but rather for their interest in the relevant subject areas and in teaching at the appropriate levels. The expression of divergent viewpoints and the searching appraisal of the test material by the members of a panel have a direct bearing upon the quality of the final paper. This panelling process brings a range of viewpoints to the preparation of the tests (from the draft stage through to the editing of the final paper), which test-writers or examiners working alone could not. Ambiguities, unwarranted assumptions, and value judgments in the questions which are likely to result from any one person's approach may thereby be recognized in the course of discussion and removed.*

The leader of the mathematics panel in 1972, David Arney, described four basic criteria used in selecting material for inclusion in the CSSE Mathematics paper: namely subject area, mathematical skills, interest and elegance, and statistical data obtained from trial testing.

Arney described the intention "to draw upon material from as diverse areas of mathematics as possible, including applying mathematics to scientific or practical data". He made some quite definite statements about the kinds of topics and skills to be tested, and made the following comments about interest and elegance.

> *It is considered important that the material be as interesting as possible. Unfortunately many students are not fond of mathematics and so the interest sought may well apply to a minority. Classification of an item or unit as interesting or elegant is usually a very subjective judgment although there is considerable agreement on boring questions.*

The leader of the science panel, Peter Thomson, described the criteria for constructing the science test in terms of subject matter, scientific skills, interest and relevance, and trial test data.

> *One aim was for the stimulus material to be scientific in nature, meaning that it can deal with any field of science and technology or it may be related to a scientific investigation of a non-scientific field. … Another aim is to present a balanced approach spanning as many subject areas as possible.*

Specific skills tested included the following:

- an understanding of the way scientists go about their work (this is sometimes referred to as the 'scientific method');
- the ability to handle scientific data in all its forms, e.g. reading tables, graphs, flow charts; following circuit diagrams; analysing scientific writing;
- the recognition of scientific and mathematical models; and
- the classification of scientific information.

Thomson gave particular emphasis to the interest and relevance of the stimulus material.

> *It is considered important that the paper does not become one which only selects those candidates who are best able to wade through an accumulation of dull, uninteresting material. Considerable thought is therefore given to providing a paper which can hold a student's interest. We would like to think that some students even enjoy the experience. Relevance goes hand in hand with interest. Apart from subject area relevance, an attempt is made to include material closely related to the problems of today's society.*

The comments of Grant, Arney and Thomson show aims and intentions quite at odds with psychometric test items discussed above. They formulate the desiderata of the cross-curricular test item writer and test constructor.

CSSE items and sub-tests aimed to be:

- broad and diverse;
- significant and meaningful;
- relevant and authentic thinking; and
- of interest to candidates and worthy of sustained examination.

CSSE tests were to be constructed on the basis of professional judgements about breadth and balance rather than a 'tightly structured specification'.

After the CSSE had evolved into the ASAT, Richard Bell did a detailed psychometric analysis of the second form of ASAT (ASAT-B) to address the following issues (Bell, 1977).

- How effective is the test as a measuring instrument?
- What does the test measure?
- What factors other than ability affect test performance?

Bell used French's *Kit of Cognitive Reference Tests* to explore the relationship between aspects of the ASAT and the relationship of ASAT to other tests. Bell identified the factors of symbolic comprehension, visual manipulation, mathematics, common reasoning, verbal reasoning, verbal knowledge, inductive reasoning, and visual/verbal interpretation in the ASAT-B. Overall, he concluded that ASAT-B could be generally described in terms of verbal and quantitative reasoning according to Figure 6.

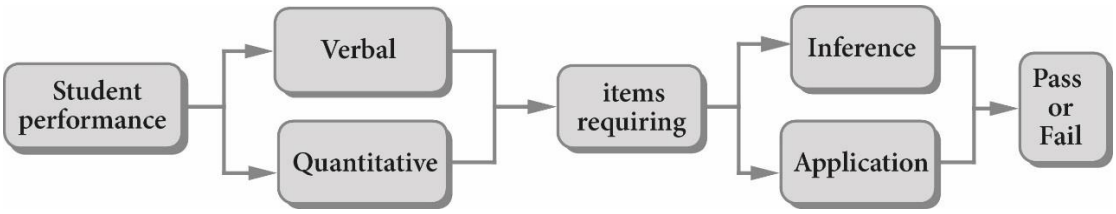


Figure 6. ASAT-B in terms of verbal and quantitative reasoning

After various kinds of analysis of the sub-tests, units and items of ASAT-B, Bell concluded that:

> *This study provided a comprehensive picture of the psychometric structure of an ASAT test. ASAT-B was shown to be both reliable and valid - in the sense that the test related well to measures of ability. Composition of the test items was shown to be both varied and complex, and there can be little doubt that the test relates to the complex abilities envisaged by its designers.*

Critical reading

The unit, *Sir Henry Parkes and Alfred Deakin* (see Exhibit 29), is a quite simple setup but it is an exemplary piece of CSSE Humanities. In the first passage, Henry Parkes gives an account of his own character and motivations. In the second passage Alfred Deakin offers a description of Parkes' character. The unit is no more than a comparison of the two views.

Other than the fact that it is a kind of self-justification, the significance of the Parkes' passage is not obvious, and the balance of considerations and the upshot of the Deakin passage is complicated. The first six items of the unit deal with the overall purpose of Parkes' statements, and they involve construing various details and comments. The first item deals with the overall purpose of the statements of Parkes. The second and third items deal with some of the more detailed comments in the passage.

Items 11 and 12 about the end of the Deakin passage hit the centre of the comparison in that they relate the comment in the last sentence about Parkes' qualities being his "both by nature and by act" with the earlier comment about: "Movements, gestures, inflexions, attitudes harmonised, not simply because they were intentionally adopted but because there was in him the substance of the man he dressed himself to appear", which is a very finely balanced comment on Parkes by Deakin. Item 12 asks candidates to see the implications of this linkage by understanding Deakin's nuanced view of Parkes that however self-conscious Parkes is, he is also a 'man with the bearing and dignity of greatness'.

There are 14 good and serviceable items in the Parkes unit, but it is Items 11 and 12 that do full justice to the subtlety of Deakin's description of Parkes, and these questions are item writing at its best. Subtle as it is at its best, the Sir Henry Parkes and Alfred Deakin unit is no more than a direct comparison of the statement of Parkes about himself and the comments on Parkes by Deakin.

The unit, *Views of the Rum Rebellion* (see Exhibit 30), consists of four passages giving different views of what happened when Governor William Bligh was arrested by the NSW Corp in 1808. After 14 questions about the passages, a painting that was exhibited in Sydney shortly after the arrest was presented to candidates. Two questions were asked about the attitude to Bligh implied by the painting in comparison with the claims made in the passages. Candidates were to recognise that the picture of Bligh hiding under the bed is not mentioned in any of the passages, and to infer that this is an unwarranted slur on the Governor. Another question asks what impression is given of the soldiers in the picture, and candidates are to recognise that they seem controlled and orderly rather than the "drunken and half-drunken troops" described by one of the passages.

Exhibit 29.
Sir Henry Parkes and Alfred Deakin, Humanities, 1969
Items 1-3, 6, 8-9, 11-12

This unit consists of two passages about Sir Henry Parkes, who was often called 'the father of Australian Federation'. Passage I is written by Sir Henry himself; Passage II by a contemporary of Parkes.

PASSAGE I

I believe myself to be a proud, but thoroughly unselfish man, with a fervent and unchanging love of my fellow-creatures. I am proud of my strength to stand alone, of my power to resist forces brought against me, of the conquests I have made by my own energy and perseverance; but I feel no pride in place or position, or in the possession of the gifts of fortune, which indeed have been few with me. I have never known what it is to feel envy of others more favoured than myself, and I have never withheld my last shilling from those who needed it more than I. The influence of these feelings will explain much in my conduct which men have misunderstood.

What, I believe, has become an ingrained feature in my public character, [is] the refusal of my very nature to seek support, approbation, or honour. I have always held that honour would lose its lustre, and public position its dignity and importance, if not voluntarily bestowed by those who had power to bestow it; if not won by the recipient on his simple merits.

I cannot deny myself the satisfaction of believing that my influence has tended to conserve the good, to eradicate the evil, to strengthen the energy and to elevate the sentiment in the national character. The growth of the colony in which I have lived and laboured has been amazing in my time. If I dare not claim any large positive share in its advancement, I have little fear of being accused by posterity of having created obstacles in its path of progress. I joined its scanty population when it was little more than a scattered settlement, and I have lived to see it entering into the noble fraternity of nations. Others will arise to examine more critically the work I have done, I can only say that it has been done with an honest purpose.

1. Parkes indicates in Passage I as a whole that his reason for writing the passage was that he

A. hoped his contemporaries would feel free to criticize his work.
B. was concerned that he had laboured mightily in vain.
C. felt a need to explain the motives for his conduct.*
D. doubted whether he had contributed anything of value to the colony.

2. In the first paragraph Parkes writes of his pride

A. at achieving an envied place in society.
B. in his toughness against great odds.*
C. in his fellow countrymen.
D. in the private wealth he has accumulated.

3. In the context of the first paragraph, Parkes' reference to others 'more favoured than myself' is a reference to other people's

A. friends.
B. abilities.
C. wealth.*
D. popularity.

6.The last sentence of Passage I suggests that Parkes was concerned above all that posterity should see him as

A. without fault.
B. a man of integrity.*
C. a successful public figure.
D. welcoming criticism.

PASSAGE II

First and foremost of course in every eye was the commanding figure of Sir Henry Parkes, than whom no actor ever more carefully posed for effect. His huge figure, slow step, deliberate glance and carefully brushed-out aureole of white hair combined to present the spectator with a picturesque whole which was not detracted from on closer acquaintance. His voice, without being musical and in spite of a slight woolliness of tone and rather affected depth, was pleasant and capable of reaching and controlling a large audience. His studied attitudes

expressed either distinguished humility or imperious command. His manner was invariably dignified, his speech slow, and his pronunciation precise, offending only by the occasional omission or misplacing of aspirates. He had always in his mind's eye his own portrait as that of a great man, and constantly adjusted himself to it. A far away expression of the eyes, intended to convey his remoteness from the earthly sphere, and often associated with melancholy treble cadences of voice in which he implied a vast and inexpressible weariness, constituted his favourite and at last his almost invariable exterior. Movements, gestures, inflexions, attitudes harmonized, not simply because they were intentionally adopted but because there was in him the substance of the man he dressed himself to appear. The real strength and depth of his capacity were such that it was always a problem with Parkes as with Disraeli where the actor, posture-maker and would be sphinx ended or where the actual man underneath began. He had both by nature and by act the manner of the sage and a statesman.

8. Which one of the following possible comments on Parkes is not supported by the evidence of Passage II?

A Consummate actor.
B Cynical manipulator.*
C Wise statesman.
D Dignified leader.

9. Sentences 1 and 2 of Passage II imply that in the first place, Parkes created an impression on others by his

A physical presence.*
B false attitudes.
C outstanding intellect.
D manner of dressing.

11. The last sentence of Passage II suggests that Parkes' qualities were his 'both by nature and by act'.

Which one of the following sentences from this passage refers most definitely to the same idea?

A 'His studied attitudes expressed either distinguished humility or imperious command.'

B 'He had always in his mind's eye his own portrait as that of a great man, and constantly adjusted himself to it.'

C 'A far-away expression of the eyes, intended to convey his remoteness from the earthly sphere, and often associated with melancholy treble cadences of voice in which he implied a vast and inexpressible weariness, constituted his favourite and at last his almost invariable exterior.'

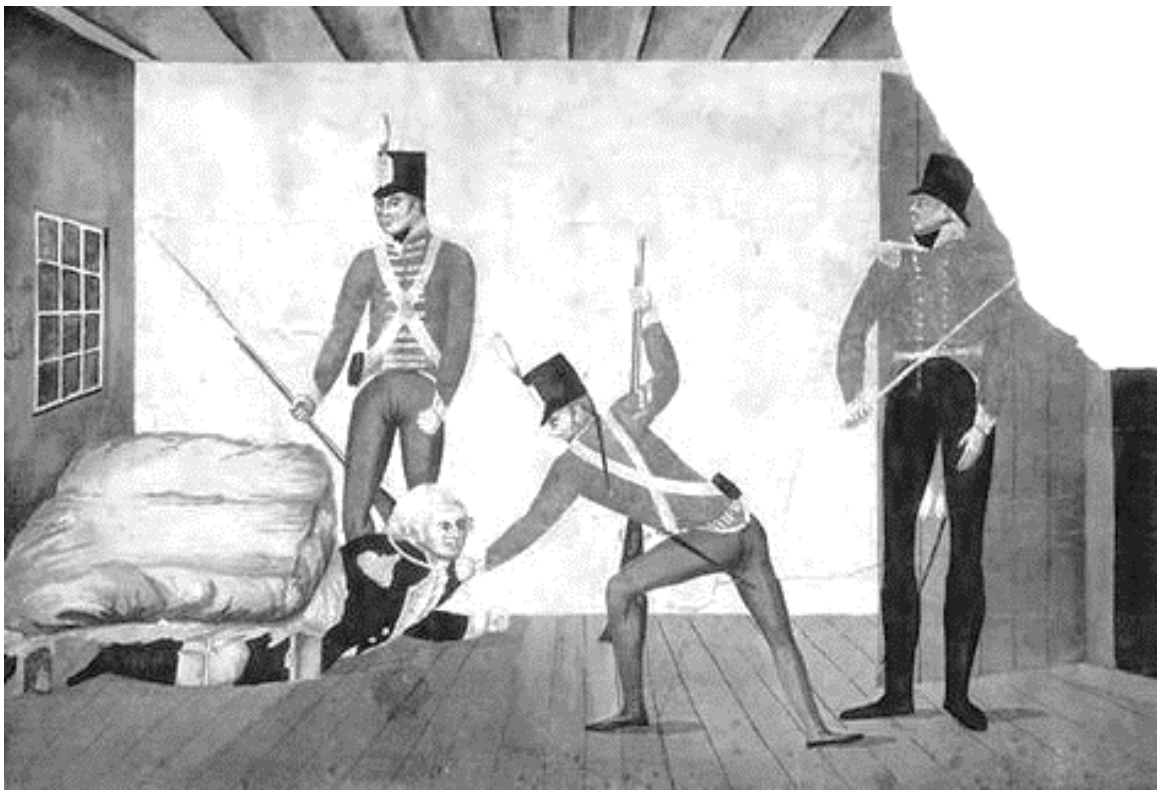
D 'Movements, gestures, inflexions, attitudes harmonized, not simply because they were intentionally adopted but because there was in him the substance of the man he dressed himself to appear.'*

12. If the author of Passage II depicts him correctly, which one of the following descriptions is best applied to Parkes?

A A man able to lead others by quiet example.
B A cunning dominating leader with an iron will.
C A man with the bearing and dignity of greatness.
D A posturer pretending to wisdom and statesmanship.

Exhibit 30.
The Rum Rebellion, Humanities 1971
Items 15-17

The painting below was a fifth piece of stimulus offered after 4 different written accounts of what happened when Governor Bligh was arrested by members of the NSW Corp during the Rum Rebellion of 1808. Candidates were asked to consider which of the accounts of the incident was closest to the impression given of the soldiers in the painting.



15. Which one of the following most accurately states the relationship between the situation presented in the painting and Passages III and IV? The situation presented in the painting

A. confirms the information given in both Passage III and Passage IV.
B. confirms the information given in Passage III but not that given in Passage IV.
C. confirms the information given in Passage IV but not that given in Passage III.
D. does not confirm the information given in either Passage II or Passage IV.

16. Which one of the following comments from the passages on the NSW Corps is the most accurate description of them as they are shown in the painting?

A. Passage I, 'the intrusion of these drunken and half-drunken troops'.
B. Passage II ,'The NSW Corps advanced under arms with fixed bayonets'.
C. Passage IV ,'a cordon of yelling Rum Corps 'bandit''.
D. Passage IV ,'a large posse of brutal creatures with muskets charged and bayonets fixed'.

17. The attitude to Bligh indicated by this painting is most like that of the author of

A. Passage I.
B. Passage II.
C. Passage III.
D. Passage IV.

7

The evolution of writing assessment at ACER

As with the other aspects of test development discussed here, the CSSE was a great stimulus to the development of thinking about the assessment of writing ability at ACER. There was a good deal of scepticism about the reliability of tests of writing ability in some assessment bodies in the 1950s and 1960s, and there was a preference for indirect tests of writing ability using multiple choice tests of language knowledge in the United States (Godshalk, Swineford & Coffman, 1966). The SAT did not include any writing in the Test of Standard Written English until 2005. On the other hand, ACER was committed from the inception of the CSSE to having a comprehensive 2 hour assessment of writing ability in the battery. As history shows, while ACER has been the major proponent of multiple-choice testing in Australia, ACER has for 50 years placed marked emphasis on 'performance assessment' through writing tests in cross-curricular batteries. Unlike most testing agencies in the United States, ACER did not have to discover 'performance assessment' in the 1990s.

The CSSE written expression test

The initial approach taken to writing assessment was outlined in the CSSE objectives of 1967 presented in Appendix 2. Writing ability was assumed to be a complex of different skills and abilities, and the test was conceived of as 'sampling the skills and abilities'. The first part of the test of 1964 was an indirect test of writing ability using a set of short answer questions about 'the specifics' of:

- grammar;
- punctuation;
- word knowledge;
- word discrimination;
- 'correctness, felicity of expression, suitability of style to purpose'; and
- the arranging and grouping of ideas.

The second part of the test focused on 'integrated skills and abilities'. In this section there were three different writing tasks, two short and one longer. The three writing tasks were different kinds of composition that aimed to test a 'large sample' of skills. Candidates had no choice of topic, but the topics were fairly general rather than specialised.

In the shorter tasks candidates were to do writing exercises like those required for subjects other than English. They were exercises such as the writing of a summary, an outline or notes. There were transactional tasks such as writing a letter, exercises such as translation of pictures or

diagrams into words, and expressive tasks such as producing a description. The longer compositions were 'creative writing' tasks such as writing imaginary or actual narratives. The first of the CSSE writing tests on the morning of Wednesday 7 October 1964 is outlined in Table 12.

Table 12. CSSE Test of Written Expression, 1964

Specific skills and abilities			
Test 1	English usage (grammar, punctuation, use of capitals)	32 items	15 minutes
Test 2	Vocabulary (discrimination)	20 items	15 minutes
Test 3	Vocabulary (opposites)	20 items	15 minutes
Test 4	Sentence sequences	20 items	15 minutes
Integrated skills and abilities			
Test 4	Short composition Synthesise ideas about newspapers	short straight-forward	15 minutes
Test 5	Short composition Describe a circus scene	description	15 minutes
Test 6	Longer original composition Narrative about advertisement in personal column	original and expressive narrative	45 minutes

The criteria for assessment

The criteria for the assessment of writing were defined in terms of purpose "since this determines the criteria to be used in assessing the extent to which he has been successful".

> *In the Written Expression paper, each topic set determines the purpose the writer (candidate) is required to fulfil through his composition. In some cases this purpose is rigidly defined; in others, some degree of freedom is given.*

The marking was a single holistic score rather than an analytical scoring of a set of criteria. It was intended that in each essay "the examiners are concerned with the *overall impression*, that is, with the demonstration of integrated skills appropriate to specific purposes". The overall assessment could involve consideration of some or all of the following specific criteria:

- content,
- clarity,
- sequence,
- unity, and
- style.

The directions to markers were focused on particular tasks, and as such they "may require consideration of additional criteria and these will be incorporated into the essay marking guides as necessary". In 1964, the marking scale was A to E with a plus or minus on each grade to give a total of 15 possible marks. Each question was marked by a different marker, and there was no second marking.

This initial design evolved so that by 1968 the 'specific skills' tasks were removed, and candidates wrote the following four pieces that were supported by stimulus material:

Test 1: When the machines cease to operate 30 minutes
Test 2: When you were left out 25 minutes
Test 3: A letter to the editor on junk and rubbish 25 minutes
Test 4: A story based on specified characters in a setting 40 minutes

In 1969, candidates wrote the following four pieces in 30 minutes each.

Test 1 *Sand – grit – dust.*
Between toes, fingers – under nails – in cupboards and shelves.
Draw on your experience to create an impression of sand, of grit, of dust.

Test 2 Candidates were invited to "write about what towns or cities mean to you". They were offered the lyrics of two pop songs (*Little Boxes* and *Downtown*) and John Brack's painting entitled, *Collins St at 5.00,* as stimulus.

Test 3 This test offered some comments from a newspaper about television violence, and invited candidates to: "Express your reaction to and opinions of the issues presented in the newspaper report as clearly and coherently as you can".

Test 4 This test was based on a photograph of a man who seems to be a Mexican peasant. Candidates were requested to imagine they have met and talked to the man, and they were asked to describe how he looks and acted.

In 1972, there were the following four tasks of 30 minutes each.

Test 1 The first tasks offered the following comments and some images of faces.

Sometime we are not what we seem.

Sometimes we are forced,
sometimes we decide,
to take on a character, a personality
which is not our own.

Sometimes the mask slips, and we stand revealed.

Test 2 Some mysterious and atmospheric photos were offered as "a starting point for a piece of writing in any form you want".

Test 3 This task was based on a picture of an old man and woman, and candidates were invited to imagine they had interviewed the people and were to write a biography of them.

Test 4 Candidates were offered two poems as stimulus and asked to "describe as vividly as you can a place that means a lot to you".

The test continued to evolve along these lines. The last year of the test was 1973, and in that year there were still four tasks of 30 minutes each.

Test 1 The first task offered some written and visual stimulus and invited candidates to use one or more of the pieces of material on the page as the starting point for a piece writing. Candidates were told "You need not know how your writing might finish before you begin".

Test 2 This task was a mysterious image of what seems to be the inside of a well and candidates were invited to 'use it as the basis of a piece of writing in any form you like'.

Test 3 This task offered some written and visual stimulus and directed candidates to write about '*Barriers*'.

Test 4 This task offered the following prompt and a set of photographs of various younger and older people.

Write about the things
which made you happy – or unhappy-
when you were younger.

The commitment to a two hour writing test was a key aspect of the CSSE battery. As we will see below, ACER was never again able to get a two hour, four-part writing test in any other battery, but it has been able to sustain two 25 minute writing tasks in CSTP, and an hour long writing tests of two parts in STAT and GAMSAT. The ASAT/AST, as we see below, is a special case having one piece of writing of 600 words to be completed in 150 minutes.

The prominence given to the assessment of writing ability is a distinctive characteristic of cross-curricular testing at ACER. When most of the world's important generic skills tests were totally multiple choice, there was (and still is) a consistent emphasis on the direct assessment of writing ability at ACER. The distinctive characteristics of the assessment of writing ability as ACER can be seen in the evolution of the CSSE writing test sketched above.

As the CSSE test evolved, the writing tasks became less particular or specifically differentiated, and the form of response was no longer specified. The topics became what might be called 'free writing exercises' rather than more specific language production exercises. Candidates were invited to express themselves and write what they liked. They were encouraged to a freedom of expression, and by the last years of the CSSE they were given the extraordinary advice that: "You need not to know how your writing might finish before you begin".

While the functional writing tasks of the first few years turned into expressive prompts, there was an even greater change to the style of marking the CSSE writing test over the 10 years.

The revision of the CSSE objectives of 1972

The Chief Examiner of CSSE from 1964 to 1969 produced a text called *Marking English Compositions* (Gosling, 1966) in which he explained the rather schematic approach taken to the marking of CSSE essays under his guidance. As the view of writing assessment evolved in the first few years of the CSSE, the objectives for the Written Expression paper published in1967 were revised in 1972. The revised objectives showed a fundamental shift in the way writing was seen and assessed (see Appendix 3).

The multiple-choice testing of grammar and word usage was discontinued and the test was "totally given over to eliciting pieces of original writing". The revised objectives offered to explain the changes in the tests by explaining the way the essays were marked.

Types of writing

According to the revised objectives, the design of the test was no longer based on 'types of writing':

> *We would not therefore now wish to suggest that a candidate should have achieved a degree of proficiency in any particular types of writing ability … Sometimes the wording of a particular topic will suggest or imply an appropriate form or mode for response, but if a pupil ignores or fails to see such implications he or she is not penalised.*

The revised objectives were concerned with a general writing ability rather than the ability to write specific kinds of things.

Topics

According to the revised objectives, one of the chief criteria for choosing topics was that they "worked well in trial testing at the appropriate grade level" (see Appendix 3). Experience had shown that some candidates performed differently on different topics whether the topics were of the same kind or of different kinds. It was not seen as necessary to require candidates to do different kinds of writing to assess their general writing ability, but it was desirable to have a number of writing samples.

The topics were chosen to "provide as wide a range as possible of stimulus material and themes - the ways in which a theme might be taken up are left for the student to review and choose from."

The aim was to give candidates "a chance to write, and to choose their own mode of discourse for the topic. A particular topic will produce essays in many modes: a single task will bring forth arguments, conversations, descriptions and stories in prose or poetry, and sometimes dramatic form."

The intention of the topics was explained as follows (See Appendix 3).

> *The main purpose of each topic is to get pupils to write, as well as they possibly can, especially since Written Expression is the only paper of the four which permits the candidates to respond in original and extended forms. We have reason to believe that the removal of all rigid definitions of how they shall write has encouraged this.*

A criticism of this approach was anticipated and rejected on the basis of the work candidates produced:

> *At first sight, such a testing technique might seem to place the very imaginative child at an advantage, but a survey of the essays which receive high marks reveals that a large proportion of these are simple, straightforward expressions of opinion or statements of fact, stories or descriptive prose, alongside the highly imaginative, creative pieces of work.*

As part of the removal of 'rigid definitions', markers were asked "not to exercise too strict a definition of relevance when they come to mark the essays". It was claimed that material elicited by the test "suggests that pupils operate well in this situation".

The marking

The description of the marking given in the revised objectives signals a clear shift in the locus of control:

> *Since 1967, the weight of responsibility for the setting of standards of 'good writing' has shifted from the Chief Examiner to individual markers. Each person is concerned, in making judgments, with the 'overall impression' as before; that is, with the demonstration of those integrated skills mentioned above. No longer, however, does the Chief Examiner define, or set limits to, the acceptability or appropriateness of particular skills to specific purposes. No marker is now required to give consideration to any specific criteria 'appropriate' to the task set, unless they are his own criteria.* (See Appendix 3)

The team of 200 markers was asked to use "the criteria which they would normally use, as teachers, in deciding one piece of writing was better than another". This statement was justified on the basis that the markers are experienced and qualified, and eight different markers assessed the work of each candidate:

> *But since all markers are trained, qualified and experienced teachers of English, some account of the specific and integrated skills will be taken by each marker. Some markers are likely to include specifics such as grammar, punctuation and word usage high in their sets of criteria;*

> *others are likely to make much less of such specifics and concentrate on integrated skills, such as narrative power, ability to write a reasoned argument or expression which has a high degree of originality in form or subject matter. Assigning eight markers to each candidate's work ensures that a cross-section of opinion among English teachers as to what constitutes 'good' writing is brought to bear on each paper.*

This approach to the marking process is based on the view there is no 'true score' and that the same matters are at issue in all the writing tasks. The candidate's assessment is the aggregated score from eight markers. The marking process was not based on detailed specification:

> *However, we wish to point out that essay marking guides, prepared by a chief examiner or panel, are no longer used. Reliable assessment can be achieved in essay examinations without this sort of external determination of purposes and criteria if the marking team is carefully chosen and the distribution of scores is controlled statistically.*

The marking was undertaken in a 'professional spirit':

> *We rely on the training, experience and professional integrity of each marker to make his or her mark valid. Investigations of the reliability of the marking are made from time to time; they give us reason to believe that candidate's writing is treated sympathetically and fairly and that total scores are a reliable and valid reflection of the criteria currently used by Australian teachers of English in the assessment of writing ability.*

Producing a spread of scores

The approach to marking described in the revised objectives is normative rather than criterion referenced:

> *The markers are asked to award grades in a distribution which very roughly accords with the normal curve: checks encourage us to believe that this technique, in conjunction with all the other setting and marking aspects of the scheme outlined above, allows a free expression of ideas and opinions in candidates' writing, encourages the demonstration of integrated writing skills, and ensures that the total scores achieved for four pieces of writing are a reliable consensus of opinion about the worth of the writing thus summoned from the candidates.*

The standards applied in this approach are temporally and culturally situated rather than absolute.

An overview of the evolution of CSSE written expression test

As in the initial years, the later years of the CSSE written expression test did not offer candidates choices between tasks, but the more specific tasks were removed, and the larger tasks became very general and unconstrained:

> *Options are still not given, and care is still taken to avoid topics where specialized or superior knowledge will confer an advantage on particular pupils.* (See Appendix 3)

The topics were very general and open, and candidates could choose their own mode of discourse for the topic.

The program did not attempt to define 'good writing' with criteria, described levels of performance or exemplar scripts. Markers were to make a professional judgement in a case by case fashion on the basis of the criteria they had developed through their work as teachers. Markers were directed to produce a range of marks that "very roughly accords with the normal curve". The revised objectives signalled a change in the locus of control in the marking, and an intention to cultivate a professional marking culture.

When the CSSE test was discontinued in 1973, the same approach to writing assessment was used in the CSTP, and it has continued to influence the approach to the assessment of writing in ASAT/AST, STAT and GAMSAT to the present time.

CSTP written expression test

The **Cooperative Scholarship Testing Program** (CSTP) developed alongside and reflected the structure and approach of the CSSE. There were two writing tests of 25 minutes for each of the three levels of CSTP testing: Level 1 was primary into secondary school, Level 2 was junior secondary into middle school, and Level 3 was middle into senior secondary school. There were more narrative topics at Level 1 and more discursive topics at Level 3, but the tasks of different levels were not sharply distinguished. There were tasks that were used at all levels of CSTP on occasion.

There was no particular pattern for the tasks at a level of CSTP in a year. One might be more specific and the other might be more open, but it would be possible for a candidate to write two stories or two more or less reflective/expressive pieces to most pairs of tests at a level in a year. Rather than being tests of form or function, the tests were open and expressive prompts like the later years of CSSE that offered candidates opportunities to write.

Some CSTP writing prompts

There were some rather specific story prompts used in CSTP, such as the following item from Level 1, 1972.

Read this note:


URGENT !

Dear D.

Meet me at the south door of the Forest Hall tonight, 10 p.m. Interesting developments concerning Mr P.

B.R.

P.S. Secrecy is essential at this stage!

Imagine you are either ‘D’ or ‘B.R.’ and write a story in which this note plays a part.

There were some rather more complex narrative prompts as indicated in this 1974 Level 2 item.

Read this passage:

The wind blew steadily out of the east, and the night air was bitter. Nothing moved across the land. Under the high cold stars lay an empty, frozen plain.

Over everything lay the silence of death.

Use the passage as a starting point for your thoughts about a piece of writing. You may write in any form you wish; for example, story, poem, or conversation. Do not write the passage out again.

There were no transactional tasks in CSTP. The emphasis was on creative or expressive writing. The crucial characteristic was the invitation that: ‘you may write in any form or style you like; for example, story, conversation, poem or description’. The prompts were very open:

Describe the sea as you like it, or write a story in which the seas moods are important.

The aim was to stimulate rather than constrain what candidates did. The direction to candidates was: ‘you can write in any way you like’.

> Use this passage as a help to your thoughts about a piece of writing of your own. You may write in any form you wish, for example a story, a poem, or a conversation.

There were some purely visual prompts as given in this 1975 Level 2 item.

Write about the man in this picture.

There were visual and text prompts of a more descriptive or narrative kind, such as the following Level 2 item from 1974.

> In this test you are asked to do a piece of writing in which fire—big or small, friendly or dangerous—plays some part. You may use the photograph below as a starting point, or you may follow your own ideas.
>
> You may write in any form or style you like; for example, story, conversation, poem or description.

Some stimulus was evocative and poetic, as demonstrated in this Level 1 item from 1979.

> Read the following:
>
> *Blood punches through*
> *every vein*
> *As lightning strips the*
> *window pane*
>
> *Under its flashing*
> *whip, a white*
> *Village leaps to light*
>
> *My mum hates thunder*
> *She plugs her ears with*
> *a towel*
> *And lies on the settee*
> *As though someone were*
> *coming to get her*
>
> *A whole week of cold April rain*
> *has swum all the feathery leaves.*
>
> *I see my face in a puddle.*
> *In the water I am crying.*
>
> *My raincoat smells like a fish.*
> *My hat has caught a creek.*
>
> Write about 'The Storm'. You may write in any form you wish: for example, a letter, a story, or a conversation. (Do NOT write out any of the poems above again).

Some stimulus was enigmatic and prompted reflection, as given in this Level 2 item from 1979.

> Read the following words carefully, and use some part of what you read as the starting point for a piece of writing.
>
> You can write in any way you like. You do not need to write out the quotations again.
>
> *I would think*
> *Until I found*
> *something*
> *I can never find;*
>
> *Something*
> *Lying*
> *On the ground*
> *In the bottom*
> *of my mind*
>
> *A Boy's Head*
> *'I feel that there's a ball*
> *in my mind. . . this ball turns*
> *every now and then and that's*
> *what makes me change moods.*
> *I have my learning mood, my*
> *good looks mood, my happy*
> *mood, my loose end mood and*
> *my grumpy mood.*
>
> *In it there is a space-ship*
> *and a project*
> *for doing away with piano lessons.*
>
> *There is a river*
> *that flows upwards.*
>
> *There is a multiplication table.*
> *There is anti-matter.*

The rationale for the CSSE and CSTP writing prompts of the 1970s

There was a clear change from a mixture of transactional and original writing to predominantly expressive writing in the 1970s in both CSSE and CSTP. This change was in tune with developments in the teaching of writing in the 1960s and 1970s. In 1968, James Moffett produced an influential argument about the diverse and creative nature of writing. In 1970, James Britton (1970; 1975) identified expressive, transactional, poetic purposes for writing, and argued for the fundamental nature of expressive writing. There was an increasing desire in this period to engage and motivate students to write rather than constrain them with rules and requirements.

The pressure of marking large numbers of CSSE scripts in a short amount of time stripped down or honed the CSSE marking process. Marking was seen as an exercise in tacit professional judgement made at speed rather than a constrained and deliberate process of ratiocination based on the application of specified criteria and defined standards. Double marking (and discrepancy procedures) were seen as a necessary safeguard in such a streamlined marking process.

The CSSE marking had begun with a conventional assumption that there is a true score for a writing script, and, for want of any other way of determining the true score, it was taken to be the score that the Chief Examiner would or did give to a script. The CSSE and CSTP evolved into the view that there is no 'true score' for a writing script, rather that there are plausible and implausible scores. On the six point scale that came to be used for marking in CSSE and CSTP, a range of two score points (a difference of no more than one between two scores) was seen as a plausible range.

This recognition of the limits of precision in the scoring of writing scripts required multiple tasks, double marking and discrepancy marking rather than the pursuit of a true score.

The evolution of CSTP

CSTP followed the CSSE pattern into the 1980s. Towards the end of the 1980s there were some signs of evolution in CSTP. A set of three criteria were identified as the ACER criteria for the assessment of writing.

THE ACER CRITERIA FOR ASSESSMENT OF WRITTEN EXPRESSION

In assessing written expression markers at ACER will consider the following issues:

THOUGHT AND CONTENT
(the quality of what is said in the piece of writing)

- what is made of and developed from the task
- the kinds of thoughts and feelings offered in response to the task

STRUCTURE AND ORGANISATION
(the quality of the structure and organisation developed to present what is said in the writing)

- the shape and form of the piece
- the sequence and cohesion of the piece

EXPRESSION, STYLE AND MECHANICS

(the quality of the language used to organise and present what is said)

- the effectiveness and appropriateness of the language
- the expressiveness and fluency of the language
- the control of the mechanics of English

The following comments that reflect the revised objectives for the CSSE were offered to the markers of CSTP in 2014.

CSTP MARKER BRIEFING

The CSTP WRITING TEST is an assessment of writing ability in which candidates are required to respond to two topics.

In assessing the CSTP WRITING TEST markers are to make a holistic impression rating using their own judgement as to what constitutes ‘writing ability’.

Markers work within the broad parameters outlined by the Criteria for the Assessment of Written Expression, but markers are not to assume the stated criteria mean they are expected to do this or that. They are expected to use whatever hunches, intuitions, or preferences they normally use in deciding that one piece of writing is better than another.

There is no attempt to fix standards for each of the categories A to E by means of sample essays; and there is no pass or fail mark at any point on the scale, i.e. D does not mean a fail essay or C a pass. D simply indicates that the marker thinks an essay is `not as good' as an essay s/he has awarded a C.

The degrees to which the quality of the ideas, the structure and organisation, and the expression found in writing are to be taken into account in marking cannot be specified. The weighting will change for different pieces depending on the strengths and weaknesses displayed by particular candidates.

The essays are to be graded on the basis of the total impression given by the piece. Markers are to take into account each criteria and should not be side-tracked by the use of a single issue such as spelling or irrelevancy.

Fundamentally in the CSTP WRITING TEST, the topic offers candidates opportunities for writing. Candidates must indicate that they are responding to the prompt rather than rehearsing a piece they have written before, but the assessment is concerned with the quality of what has been written rather than whether it is a 'correct' response to a particular set of instructions or a topic.

All markers have personal preferences for, but more often against, certain points of view and subjects or kinds of writing and thinking. It is important that markers recognise these and refer to the Supervisor any essay which they feel unable to assess without irritation or prejudice.

CSTP MARKING PROCEDURES

Without distorting the view they have of the work they read, Markers are to approximate the following distribution for each folder of 100 candidates:

THE GRADE SCALE

A+ - to be used for the outstanding essay

A - approximately 10% of essays

B - approximately 24% of essays

C - approximately 32% of essays

D - approximately 24% of essays

E - approximately 10% of essays

Z - insufficient to assess

X - absent

A tally of more than 14 scripts outside this distribution will mean that the group of marks will be removed and another set of marks will be obtained. While markers are to aim to spread the marks, they must not record and tally the number of grades they produce while marking.

The following comments are offered to CSTP markers about the criteria for assessment.

COMMENTS ON THE ACER MARKING CRITERIA

This assessment is a matter of individual or case by case judgements. The great challenge is to be sensitive enough to make sense of and assess what candidates write, rather than impose a scheme or expectation on the candidates. The criteria for assessment are the same for different kinds of topics and they can cover different kinds of responses or mixed responses from candidates. Writing is assessed here as a kind of thinking that involves both generating ideas and linguistic production and organisation. The overall aim in marking is to keep consideration of the quality of the thinking in balance with consideration of the quality of the writing at sentence and whole text levels.

We interpret topics liberally. They are springboards rather than straight-jackets. We do not assume that certain kinds of topics require certain kinds of responses, but we look at possibly pre-prepared essays very sceptically. (Consult the marking supervisor?)

There may be significant virtues in papers that show uneven control of language. Do not apply a negative model. Strong papers will often be uneven. Not making mistakes is not a major virtue.

We should not be overwhelmed by the showy and meretricious or the melodramatically emotive. The most difficult papers to see fairly are the quiet and substantial ones.

We treat what candidates say with a pluralistic tolerance. We are careful not to favour our own predilections and preferences.

Distinctive features of CSTP marking of writing

It will help to see the distinctive nature of the CSTP approach to marking writing by contrasting it with the analytical single marking used in the **National Assessment Program – Literacy and Numeracy (**NAPLAN**)** that was introduced across Australia in 2008.

The NAPLAN prompts are quite open, and could be used as CSTP prompts, but the assumptions on which the marking of the scripts is based contrasts markedly with the CSSE/CSTP tradition. The NAPLAN marking is a single analytical score using the 10 criteria shown in Table 13 that are added for a total score.

Table 13. National Assessment Program – Literacy and Numeracy (NAPLAN) Score Criteria

Audience	0–6	13%
Text structure	0–4	8%
Ideas	0–5	10%
Character and setting	0–4	8%
Vocabulary	0–5	10%
Cohesion	0–4	8%
Paragraphing	0–2	6%
Sentence structure	0–6	13%
Punctuation	0–5	10%
Spelling	0–6	13%

The instructions to NAPLAN candidates give a model for the answer, and that model is reflected in the marking criteria.

The NAPLAN Persuasive Writing Test of 2012

Everyone should learn to cook.

Do you agree?
Do you disagree?

Perhaps you can think of ideas for both sides.

Write to convince a reader of your opinion.

Start with an introduction.
- An introduction lets a reader know what you are going to write about.

Write your opinion on the topic.
- Give reasons for your opinion.
- Explain your reasons.

Finish with a conclusion.
- A conclusion sums up your reasons so that a reader is convinced of your opinion.

Remember to:
- plan your writing
- use paragraphs to organise your ideas
- write in sentences
- choose your words carefully to convince
- a reader of your opinion
- pay attention to your spelling and punctuation
- check and edit your writing so it is clear for a reader.

CSTP writing uses general and open writing tasks that challenge candidates to construct a response rather than follow a declared or assumed model. Candidates have to work out how they will develop a response rather than working out what it is they are supposed to do.

The marking of CSTP is norm rather than criterion referenced. Markers are given a certain amount of direction through the criteria for assessment, but they are not asked to apply a model of a good piece of writing or a good answer, and they are not offered examples of various levels of performance. Markers are challenged to make case by case judgements about the quality of the scripts they read.

The ASAT writing test

In 1984 the Board of Studies in Canberra asked ACER to develop a new writing test as part of the ASAT program. This writing test was developed in consultation with Board staff and teachers from ACT colleges.

The ASAT writing test is a unique writing test. It is an assessment of verbal reasoning and writing ability in which candidates are requested to respond in an argumentative mode to a broad range of stimulus material on a social or political issue. AST writing papers are an A3 sheet of newspaper articles, quotations from other texts and a cartoon or other image on a broad theme. ASAT writing was designed to be an authentic and hence valid test of the 'writing process'. ASAT writing is a 'process writing test' in that candidates have 150 minutes to write 600 words, and they are directed to write a draft and a finished copy of the piece in the time available. Only the finished draft is collected and assessed. Candidates are given the following instructions in the ASAT writing test.

INSTRUCTIONS

Read carefully the material on these two pages.

Write an essay of about 600 words, giving your point of view on the major issue raised in the material.

Give your essay a clear title.

You should assume you are writing to an intelligent adult audience.

The stimulus on the paper offers a range of material with a more or less common theme. It is described by those who develop it as a 'questionless question' in that no specific issue is offered for discussion. Quite different aspects of the theme can be selected for discussion by individual candidates, and they can develop a piece of writing in quite different ways. As a 'questionless question', the AST writing test is a challenge to generative thinking for candidates in that they have to select and define an issue for themselves. This characteristic of the task contrasts with writing tests which emphasise answering a specific question, and use the constraint of 'relevance' to the topic as a key criterion for assessment.

Many high stakes writing tests are specific and constrained tasks. The stimulus material offered for the AST writing test, on the other hand, gives candidates a good deal of scope for constructing their own response to the broad theme. The 'questionless question' challenges candidates to identify and define an issue, and to construct a point of view about that issue. The questionless writing test is in the spirit of the open writing tasks of CSSE/CSTP for senior students.

The pattern of specific and constrained tasks in many high stakes writing tests suggests that these tasks are designed to facilitate grading on the assumption that the more specific the task the more easily markers will be able to rank the scripts and to agree with each other. But the more specific a task the more likely it is to be amenable to some candidates and disagreeable to others. Where candidates do two different tasks in a writing test it is not uncommon for them to get substantially

different scores for each piece. This difference between the two pieces by the same candidate suggests that they find some topics more amenable than others, and the more specific the topic the more likely a candidate will find it amenable or disagreeable. One way of lessening this topic effect in the assessment of writing ability is to set broad and general themes, and allow candidates to decide how and in what way they want to respond to the theme. It is for this reason that the ASAT writing test is a broad and open task that invites candidates to construct a response rather than answer a question. Because it is a broad and open task, the ASAT writing test elicits generative thinking from candidates.

STAT and GAMSAT writing tests

Two writing tests were developed in the early 1990s at ACER that adapted but continued the CSSE/CSTP tradition. The **Special Tertiary Entry Test** (STAT) is used for alternate entry to tertiary institutions, and the **Graduate Australian Medical School Admission Test** (GAMSAT) is used for entry to graduate medical programs.

As they were aimed at older candidates than the CSSE/CSTP, these tests did not use the visual prompts or offer invitations to narrative. There were two tasks in each of these tests, but there is little distinction made between the tasks. The tasks in these tests were slightly distinguished on the basis that one is focussed on more impersonal or current affairs topics, and the other is focussed on more personal or informal topics. In the CSSE/CSTP spirit, candidates were not directed or expected to produce a certain kind of writing. Each task is a broad theme, and a number of comments and quotations are offered about the theme. Candidates are directed to: "Consider the following comments and develop a piece of writing in response to one or more of them". Candidates were told that:

> *Your composition will be judged on the quality of what you have to say in response to the theme, how well you organise and present your point of view, and how effectively you express yourself.*

The following are examples of the kind of themes and comments offered to GAMSAT candidates.

Topic A

Comment 1
Most crime is a result of disadvantage and injustice.

Comment 2
Respect for the law is declining because of increasing selfishness and greed.

Comment 3
A zero tolerance policing approach can reduce crime.

Comment 4
There is one law for the rich and another law for the poor.

Topic B

Comment 1
What we see as entertaining reflects what we are and also makes us what we are.

Comment 2
One person's fun is another person's work

Comment 3
We can be manipulated and exploited as much in our leisure and entertainment as in our work.

Comment 4
Eight hours work, eight hours rest and eight hours play.

STAT and GAMSAT writing tests differ from the CSSE/CSTP approach in that it is intended that the results of these tests could be used over a number of years, and it is possible that the ability of the candidates could vary from year to year. The norm-referenced nature of the CSSE/CSTP approach could not be applied to these tests.

The design of STAT and GAMSAT writing tests aimed to reconcile choice for the candidates and comparability for the markers by prescribing a theme, but giving candidates a choice about which prompt or prompts they might respond to in the set theme. The test design and the marking of the scripts allows (or challenges) candidates to respond to the theme and prompts in different ways. Candidates were expected to clearly base their response on the theme and prompts, but it was not presumed that the response had to be of this or that kind. Candidates could write quite formal arguments if they choose, they could take an informal discursive approach, or they could write personal narratives. Most candidates write a more or less personal expression of their point of view. They might move between more or less personal styles of writing, and in a few cases they write imaginative narratives.

As well as using versions of the ACER Criteria for the Assessment of Writing, these programs use described scales as the basis of the grading. In the spirit of the CSSE/CSTP, these projects were cautious about claiming to offer models of different levels of performance because such models are always in danger of misrepresenting the diversity of candidate scripts at different levels of performance. The levels of achievement for STAT and GAMSAT reflect the spirit of the CSSE/CSTP approach in that they are "emergent rather than prescriptive". The descriptions are said to "offer indicators of different levels of performance, but they are not comprehensive nor do they specify a necessary minimum". Markers are challenged to "make subtle case by case decisions within the guidelines set by the GAMSAT criteria for assessment of writing". The notion of 'case by case decisions' recommended in STAT and GAMSAT reflects the approach to marking that evolved in the CSSE.

A distinct line of continuity and development can be seen between CSSE, CSTP, STAT and GAMSAT. The evolution in the design and marking of writing tests sketched above shows:

- a preference for broad, open and unconstrained writing tasks;
- a rejection of formalist and functional approaches to writing test design;
- an encouragement of more personal, reflective and expressive rather than transactional writing;
- a challenge to candidates to choose or construct a form for a response;
- an emphasis on creative and generative rather than convergent thinking in writing;
- the use of holistic impression marking rather than analytical marking;
- double marking and discrepancy marking;
- an emphasis on case by case decision making rather than rather attempting to specify quality through specific criteria, described levels of performance and exemplar scripts;
- a recognition that the quality of a marking regime depends on the quality of the markers involved; and
- an attempt to develop a professional marking culture and an avoid heavy-handed marker 'training' and marker monitoring procedures.

This approach evolved out of the quite different approach described in *Marking English Compositions* of 1966. This different culture can be seen developing in *English and its Assessment*

(Maling Keepes, 1973) and the revised CSSE objectives for writing of 1973. This approach was used in the CSTP writing test, which in turn shaped the approach taken to writing assessment in STAT and GAMSAT.

Exhibit 31 on the *Deadulus and Icarus legend* shows the way the CSSE Humanities test developed. There were five pieces of stimulus and 13 items in this five-page unit. It begins with a prose account of the Daedulus and Icarus legend, followed by a passage about Stephan Dedalus from *The Portrait of the Artist as a Young Man* by James Joyce. This passage is followed by *The Musee des Beaux Arts* of W. H. Auden, the painting *Fall of Icarus* by Breughel and the ceramic painting *Icarus Fallen on the Field* by Arthur Boyd. None of the creators of the stimulus was named.

Exhibit 31.
The legend of Daedulus and Icarus, Humanities 1969
Items 33, 36-37

This unit consists of material connected with the legend of Daedalus and Icarus. There is an account of the legend, and extract from a novel, a poem and reproductions of two paintings.

About suffering they were never wrong,
The Old Masters; how well, they understood
Its human position; how it takes place
While someone else is eating or opening a window or just walking dully along;
How, when the aged are reverently, passionately waiting
For the miraculous birth, there always must be
Children who did not specially want it to happen, skating
On a pond at the edge of the wood:
They never forgot
That even the dreadful martyrdom must run its course
Anyhow in a corner, some untidy spot
Where the dogs go on with their doggy life and the torturer's horse
Scratches its innocent behind on a tree.
In Breughel's Icarus, for instance: how everything turns away
Quite leisurely from the disaster; the ploughman may
Have heard the splash, the forsaken cry,
But for him it was not an important failure; the sun shone
As it had to on the white legs disappearing into the green
Water; and the expensive delicate ship that must have seen
Something amazing, a boy falling out of the sky,
had somewhere to get to and sailed calmly on.

33.The aspect of suffering referred to in lines 3 and 4 of the poem is best illustrated by the

A. face of Icarus in Painting X.
B. ploughman in Painting Y.
C. legs disappearing into the sea in Painting Y.
D. black birds depicted in Painting X.

36. Which one of the following best sums up a difference in emphasis between Painting X and Painting Y which is essential to an understanding of them?

A. In Painting Y the focus of attention is on everyday activities; in Painting X on the agony of suffering.

B. In Painting Y the emphasis is on the shepherd boy looking at the sky; in Painting X on the damage resulting from Icarus' fall.

C. The style of depicting the human figures in Painting Y is natural; in Painting X the human figure is badly drawn.

D. In Painting Y the background of sea, sky and mountains is clearly presented; in Painting X there is little background.

37. From a comparison of Paintings X and Y, a justified conclusion is that

A. Painting X shows more tolerance of human suffering than Painting Y.
B. Painting X is like an enlargement of part of Painting Y.
C. Painting X shows more awareness of different reactions to a disaster than Painting Y.
D. Painting X provides a different way of looking at human suffering from that given in Painting Y.

Painting X

Painting Y

The unit is an extraordinarily rich set of stimulus. Candidates had to consider the legend from a range of views and compare the different views. Item 33 requires candidates to infer a view about suffering from a point in the poem, and then asks them to identify the idea in a detail from the images. Item 36 asks candidates to contrast the two images. Item 37 calls for an evaluation by identifying a justified conclusion about the difference between the two images.

8

From multiple-choice and written tests to short-answer items for the AST

We have seen how the CSSE evolved into the CSTP, the ASAT and the AST. The very distinctive ASAT writing test, described in the previous chapter, was developed due to concerns in the ACT about the difference between the performance of females and males on the ASAT multiple-choice questions.

The Queensland Board of Senior Secondary Studies decided to develop their own scaling test which was introduced in 1994. The Queensland Core Skills Test (QCST) clearly grows out of the CSSE/ASAT tradition, but it differs from that tradition in some significant ways. In construct terms it eschews the domain-related sub-tests of the CSSE/ASAT (it has no MST or HASS), and instead bases the test on what are called the Common Curriculum Elements shown in Table 14. The QCST also differed from the CSSE/ASAT tradition by introducing a short-answer question sub-test, a mode of response that had not been used much by ACER in the CSSE or ASAT.

In 2003, the ACT Board of Senior Secondary Studies decided to broaden the AST battery by bridging the gap between the multiple-choice questions and their very open writing test. The AST short-answer question sub-test was envisaged as one cross-curricular test that would not to be broken into MST and HASS sub-scores. The attempt would be to focus on the skills that were truly cross-curricular. It was also expected that the short-answer questions (like the writing test) would assess the kind of skills that are or can be taught in class work.

The aim of AST short-answer question was to further broaden the battery, bring the performance of males and female closer together, and hopefully increase the correlations with the different subject scores of colleges. A number of considerations were taken into account in designing the Short-answer question test. The short-answer items were to contrast with the kinds of thinking undertaken in the multiple-choice questions and the writing test, and there was to be an emphasis on open rather than closed questions. The test was to be less convergent than the multiple-choice questions, and less open or divergent than the writing test. While candidates might be required to respond to short-answer questions in writing, it was not to be a test of 'writing ability'. It was also intended that the items should try to reflect what is done in class work so as to better correlate with school assessments, and it was hoped that the short-answer question would set tasks that could make meaningful class work.

It is common to refer in assessment parlance to 'constructed response items' or 'performance assessments'. These very broad terms are not used here because they cover quite different kinds of assessments. The term 'constructed response items' usually means no more than that a test item

is not a multiple-choice question. There is a very broad spectrum of constructed response items stretching from open writing tests (such as the AST writing test) to specific and constrained responses that are no more than a word or a number. The AST short-answer test was designed to have some specific questions with brief answers, but it also included some general questions that invited page length answers. The AST short-answer test is 105 minutes long and has about 20 questions that are scored on mark ranges of between one and five score points.

Table 14. The Common Curriculum Elements of the Queensland Core Skills Test

The Common Curriculum Elements	
1. Recognising letters, words and other symbols	27. Interrelating ideas/themes/issues
2. Finding material in an indexed collection	28. Reaching a conclusion which is necessarily true provided a given set of assumptions is true
3. Recalling/remembering	29. Reaching a conclusion which is consistent with a given set of assumptions
4. Interpreting the meaning of words or other symbols	30. Inserting an intermediate between members of a series
5. Interpreting the meaning of pictures/illustrations	31. Extrapolating
6. Interpreting the meaning of tables or diagrams or maps or graphs	32. Applying strategies to trial and test ideas and procedures
7. Translating from one form to another	33. Applying a progression of steps to achieve the required answer
8. Using correct spelling, punctuation, grammar	34. Generalising from information
9. Using vocabulary appropriate to a context	35. Hypothesising
10. Summarising/condensing written text	36. Criticising
11. Compiling lists/statistics	37. Analysing
12. Recording/noting data	38. Synthesising
13. Compiling results in a tabular form	39. Judging/evaluating
14. Graphing	40. Creating/composing/devising
15. Calculating with or without calculators	41. Justifying
16. Estimating numerical magnitude	42. Perceiving patterns
17. Approximating a numerical value	43. Graphicalising
18. Substituting in formulae	44. Identifying shapes in two and three dimensions
19. Setting out/presenting/arranging/displaying	45. Searching and locating items/information
20. Structuring/organising extended written text	46. Observing systematically
21. Structuring/organising a mathematical argument	47. Gesturing
22. Explaining to others	48. Manipulating/operating/using equipment
23. Expounding a viewpoint	49. Sketching/drawing
24. Empathising	
25. Comparing, contrasting	
26. Classifying	

Towards a teachable short-answer question test

While it is useful to acquaint students with the kind of thinking involved in the types of multiple-choice question found in the AST, there is limited value to be gained from trying to teach students how to deal with the multiple-choice sub-test. Multiple-choice questions are developed by finding rich stimulus and asking quintessential questions about it. The better the test, the more the questions are quintessential and unique to a particular piece of stimulus. On the other hand, the AST writing test is designed to be teachable and teachers have continued for 30 years to support the current test design because it offers a valuable stimulus to the development of writing skills, it is worth teaching and it has a 'positive backwash'. Teachers set out to teach the skills needed to deal with the AST writing test.

With the aim of making the short-answer question test teachable, there was an effort to design question that are like those asked by the teachers in normal course work, or to ask questions that teachers might use in normal course work. With the AST writing test the intention was to have no

question, and to challenge candidates to develop the nature and terms of their own discussion. The decisions they would make in doing so are part of the assessment.

It was decided to make some of the short-answer questions formulaic rather than the kind of unique and quintessential questions typical of multiple-choice. The intention was to have some of the short-answer items use more or less the same question in different years, but to vary the stimulus that the questions were applied to. For instance, the following question was used about an image in the first test of 2004.

> *What impressions are we given by the illustration on the opposite page?*
> *And what is suggested by the drawing?*
> *Your responses will be judged on the:*
> - *accuracy of the description and understanding of the material; and*
> - *substance and quality of the interpretation offered.*

More or less the same formulation has been used with different images since 2004. Two examples of the material this question was used for are shown in Exhibit 32. It is thought that teachers can find their own material to put into such question shells.

Formulaic or archetypal short-answer questions

The aim was to make the short-answer question teachable by having a substantial number of items in each test that are formulaic or archetypal. Table 15 gives a version of the initial thinking about archetypal short-answer questions.

Table 15. Overview of formulaic short-answer question items used in the AST

KIND OF PROCESS	KIND OF QUESTION	LANGUAGE LOAD	
Thinking about the human world			
		Higher verbal	Lower verbal
Comprehension	What is said? What is meant?		
Critical thinking	How strong or convincing is this? Do you agree with this? Why?		
Interpretation	How does this seem? How do you see this? Explain your view.		
Thinking about the material world			
Data processing	What is the status of this data? What does this data show?		
Explanation	How does this happen? Why is this so?		
Problem solving	What is the problem? What is the solution?		

Table 16 gives an overview of the kinds of formulaic short-answer items currently used in the AST.

Table 16. Short-answer archetypal questions

Reading literacy	• underlining text • note taking • comparing the meanings of texts
Visual literacy	• interpreting images and diagrams
Critical thinking	• analysing and evaluating arguments • making arguments for and against propositions
Research and investigation	• analysing and understanding research procedures • seeing the significance of research questions
Data analysis and interpretation	• understanding and interpreting numerical data
Critical numeracy	• making calculations and evaluating numerical procedures
Systems reasoning and problem solving	• analysing and understanding how things work • speculating about how things might work

The evolution of the scoring model

The scoring of the short-answer question test has developed significantly over the 10 administrations of the test. In the early administrations the aim was to have a balance of partial credit scoring and qualitative grading. As the kinds of questions asked evolved, the partial credit approach that assumes a correct answer and more or less correct renditions of it proved to be less and less useful. The current aim is to have about half the test made up of global rather than specific questions, and these questions are qualitatively scored with holistic grades. There has been a move away from partial credit scoring because it was found it to be comparatively artificial and limiting.

Designing a bow wave rather than worrying about the backwash

As with the AST writing test, the short-answer question test is intended to have a positive bow wave in that it aims to have teachers teach and students learn the skills implicit in the archetypal questions. The schools are preparing students for the test, and there has been positive feedback from teachers to the Board of Studies about the test.

A survey of student reactions in 2009 elicited interestingly mixed feedback. Students have a good deal of time to complete the writing test, but they are not allowed to leave the test room until the end of the test. It is common for candidates who finish their piece of work before time to sit doing nothing until the time has expired. In 2009, the back page of the writing test was used to ask candidates, if they had spare time, to answer a number of questions about the different tests. About 35 per cent of candidates offered responses to the questions, and the responses were almost always constructive.

The first two questions asked candidates which questions or parts of the test they found easiest and most difficult. A summary of comments and responses to these questions is offered in Appendix 7. The comments showed that candidates were quite clear about the differences between the sub-tests, and most of them showed a clear preference for one or other. They distinguished the short-answer questions, the multiple-choice questions, and the writing test in clear and intelligible terms. They knew that their writing was not being assessed in the short-answer question, and they knew that they had to think differently in the short-answer questions from the multiple-choice questions. The candidates realised that some of the short-answer questions were open ended, and that they had to identify what was at issue in a question and construct a response. Not surprisingly, some

found such questions liberating and empowering, and some found them vague and irritating. This diversity of response is as it should be.

Changes in performance over 10 administrations

There has been a marked improvement in performance on the formulaic questions over the 10 versions of the test. From the markers perspective, candidates are better prepared and perform better on the visual interpretation, critical thinking, and the generalised data interpretation questions, in particular.

The test places a good deal of emphasis on visual and diagrammatic interpretation. The work of candidates as a whole has always been good on such questions, but there seems to have been a clear improvement over the years in the performance of those with less aptitude for such work. Few candidates now offer the kind of blindly and blankly literal responses that were a substantial minority in the first years of the test.

While the candidates express their personal opinions very well in the writing test, it was immediately obvious in the short-answer questions that they were much less able to undertake the more impersonal and disciplined thought and argument required for the critical thinking short-answer questions. The critical thinking questions were made more specific and concrete to assist in this area. There has been a clear improvement over the years, evidenced by a move back to more global and evaluative critical thinking questions. Now, after 10 administrations, candidates are significantly more comfortable than they were in the past with open-ended questions that ask for interpretations of data. Most of them understand that they are looking for what is significant and worth interpreting in the data presented.

The teachable model

It seems clear that the archetypical questions have been used by teachers. They can be taught as a style of thinking, and it seems that the thinking involved is appropriate for a range of subjects. Anecdotal reports and the performance of the candidates from the markers' perspective suggest that teachers are using the archetypal questions in preparing students for the test, and possibly in their other class work. The approach has certainly led to improved performance by candidates, and there is general acceptance in the system that preparing for the short-answer question test is of educational value. Each year the Board runs a workshop on scoring the short-answer question test as professional development for teachers in the system. The short-answer question has come to be the sub-test that correlates best with school assessments.

Two examples of different stimulus used with the same questions

There has been an increasing interest in teaching and testing visual interpretation skills in recent years. Mind mapping, 'graphical organisers' and other graphic modes of representing ideas and data are popular with teachers. The English curriculum in many Australian systems has come to treat visual representations as another kind of text in which students learn how to interpret. Exhibit 32 presents the AST visual interpretation stimulus of 2007.

The same question as was used in 2004 was also used for the following pair of images in 2011, shown in Exhibit 33.

Exhibit 32.
AST visual interpretation stimulus, 2007

What impressions are we given by the drawing below?

And what is suggested by the drawing?

Your responses will be judged on the:
- accuracy of the description and understanding of the material; and
- substance and quality of the interpretation offered.

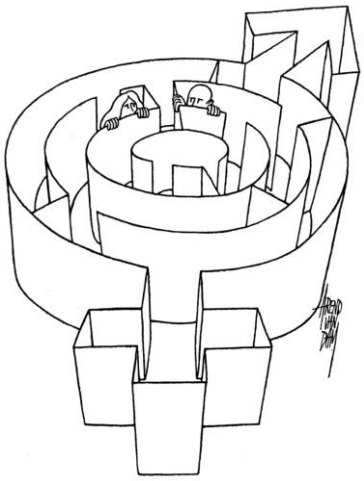


Exhibit 33.
AST visual interpretation stimulus, 2011

Image 1

A bar at the Folies-Bergère painted by Edouard Manet in Paris in 1891.

Image 2

Barmaid painted by John Brack in Melbourne in 1952.

Some archetypical short-answer questions

The stimulus presented in Exhibit 34 shows different sources of the CO_2 emissions produced in different regions of the world. There were a number of specific items to open the unit, followed by a question inviting global interpretation.

Exhibit 34.
AST interpretation stimulus, 2011

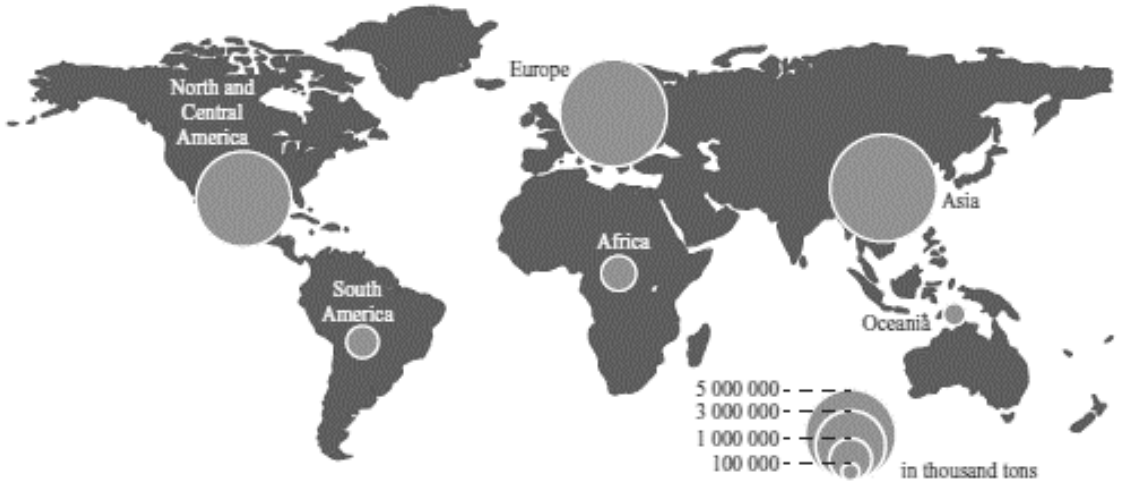


Figure 1. CO_2 emissions from energy production, industrial processes and transport

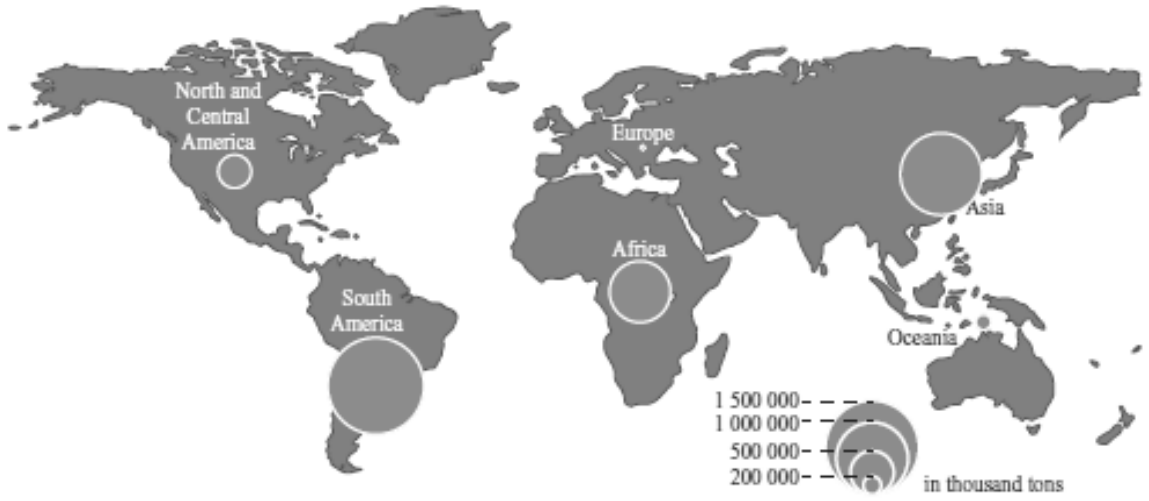


Figure 2. CO_2 emissions from land use change (cutting down native forests for agriculture, built-up areas, urbanisation, roads etc)

1. Which region shown in the figures has the highest total CO2 emissions? (1)

2. Compare and contrast Figures 1 and 2. What do the figures show about CO2 emissions? (3)

The diagrams in Exhibit 35 were identified to candidates as different views of the activity of 'problem solving' and they were offered the following question.

Exhibit 35.
Problem solving

Compare and contrast how the diagrams present 'problem solving' and what the diagrams suggest about the activity.

Your responses will be judged on the:

- accuracy of the description and understanding of the material;
- substance and quality of the interpretation offered.

(4)

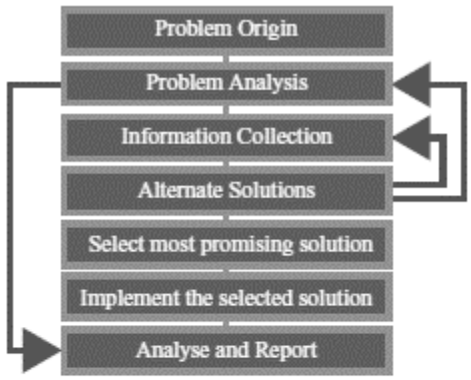


Diagram 2

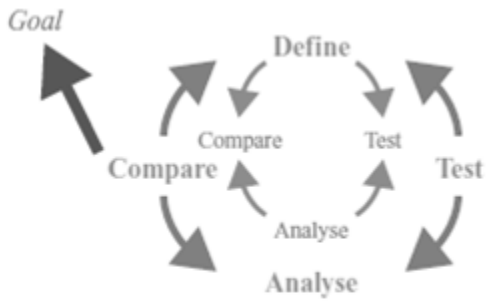


Diagram 1

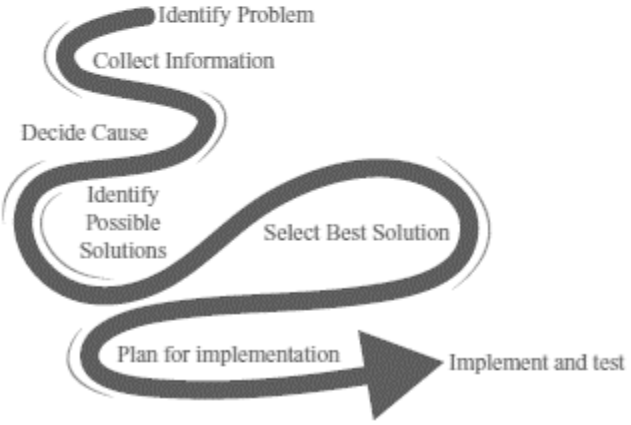


Diagram 3

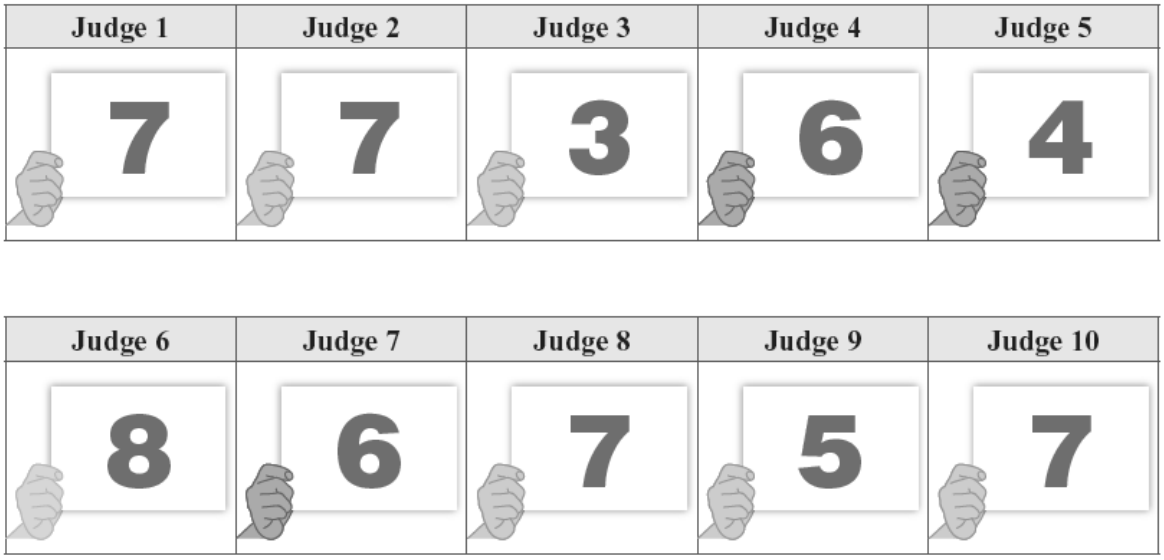
Two critical numeracy questions

Two examples of numeracy questions that promote critical thinking are presented in Exhibit 36 and Exhibit 37.

Exhibit 36.
Diving judges

In an effort to reduce mistakes and bias in awarding scores, organisers of a figure skating event use ten judges. The ten judges evaluate each competitor and award them a score out of ten.

Below is an example of the scores given to a competitor by ten judges at a recent competition.



Three methods of obtaining the competitor's overall score are proposed:

I average the ten scores

II discard a score if it is two or more above or below the most commonly occurring score, and average the remaining scores

III discard two judges' scores at random (that is, without knowing their scores), and average the remaining eight scores.

Discuss the implications of each of the three methods of calculating a total score for the competitor in the example.

(4)

Exhibit 37.
Medal tally

The table below shows the 2014 Winter Olympics medal tally for ten selected countries, ranked by three different methods.

Medal priority: The number of gold medals won by a country is counted, and the countries are ranked accordingly. For countries with equal numbers of gold medals, the number of silver medals is considered; if the number is still the same, the number of bronze medals is considered.

Total medals: The countries are ranked according to the total number of medals won.

Points system: Different medals are allocated a value (3 points for gold, 2 points for silver and 1 point for bronze). The points are added to give the total score for a country, and the countries are then ranked according to the total score.

	Rank by Medal priority	Gold	Silver	Bronze	Rank by Total medals	Rank by Points system	
1	Russia	13	11	9	Russia	Russia	1
2	Norway	11	5	10	United States	Canada	2
3	Canada	10	10	5	Norway	Norway	3
4	United States	9	7	12	Canada	United States	4
5	Netherlands	8	7	9	Netherlands	Netherlands	5
6	Germany	8	6	5	Germany	Germany	6
7	Switzerland	6	3	2	Austria	Austria	7
8	Belarus	5	0	1	France	France	8
9	Austria	4	8	5	Switzerland	Switzerland	9
10	France	4	4	7	Belarus	Belarus	10

Question 1.
When the countries are ranked using the **Medal priority** method, which countries would have their ranking determined by the number of silver medals? (1)

.....

Question 2.
Which country would benefit most by ranking using the **Points system** rather than the **Total medals** method? (1)

.....

Question 3.
Consider all three ranking methods. Which method do you believe is best? Explain why you believe this method is better than the others. (3)

.....

9

The art of cross-curricular item writing: CSSE Humanities

It has been argued here that the intentions of the CSSE led to a distinctive approach to the writing of multiple-choice test items that contrasted markedly with the usual psychometric test items. Rather than discrete items of a specific kind, CSSE Humanities was made up of sets of related items dealing with the stimulus of a unit. Units were large in scale and they were often diverse in stimulus and multi-modal. The test of 1969 had a unit on the *Battle of the Somme* with a map, five pieces of written stimulus, and some numerical material. There were 26 items constituting 27 per cent of the test on this one topic. The final CSSE Humanities of 1973 had six large units, each with multiple pieces of stimulus, totalling 86 items. A unit would usually have a theme, but the theme could be developed in different ways. Units often involved comparison of related and contrasting material. Exhibit 38 is part of the *Battle of the Somme* unit and involves critical analysis and comparison of five different views of the battle and the role General Haig, Commander-in-Chief of British forces. In this part of the unit, candidates had to recognise the different purposes and view of the writers, and assess the different explanations of what happened in the battle. The debate about whether the battle was an "overwhelming victory" was to be related to the subsequent information about casualties on each side.

There are inevitably some regularities in the questions posed in CSSE Humanities. In different years and different units the obvious and central questions about the main ideas or the implications and significance of ideas in a text were asked, but the key feature of the Humanities items is that the questions are not formulaic or algorithmic. CSSE units begin with some stimulating material about an issue, and then the aim was to find the questions of unique significance for that material. Rich stimulus material is the basis of CSSE style of multiple-choice testing.

Reviewing the 10 years of CSSE Humanities tests shows them to offer a surprising reflection of the contemporary zeitgeist. Many CSSE Humanities units dealt with the political and social issues of the time, such as 'the generation gap', the tension between India and China, and the Vietnam War. There is even a unit about The Beatles when they were just going into their hippy phase and the Maharishi was their guru. It was the 1960s. There was a certain moral undertone to some of the socio-political material that expected young people to be engaged and concerned about social and political issues. Some units dealt with controversial issues such as political protest, and some units focussed on humour, irony and even sarcasm and insult.

The items show a particular interest in comparing different pieces of stimulus material, and sometimes the comparisons involved unstated criteria and complex tacit or intuitive comparisons.

In some cases, the matching procedures involved comparing written texts with images, such as the matching of the cubist nude with the quotation from *The Wasteland*, shown in Exhibit 1. There are many other Humanities units and items that ask candidates to think intuitively and laterally rather than logico-deductively.

There is a significant emphasis in the Humanities items on aesthetic and design issues, such as the Poster unit, Exhibit 9. The items that involved visual interpretation were some of the most adventurous in the test.

Historical units often used primary sources, or took a historiographical approach and focussed on differences between historians, as in Exhibit 6, about the *Norman Conquest*, or the *Battle of the Somme* unit (see Exhibit 38). There was a certain British feel to a substantial amount of the historical interpretation in the tests, but there were units on Australian history (Exhibits 14 and 15), and there was also a certain emphasis on Asian issues and culture.

While there was some information processing material in the CSSE Humanities (often with a geographical flavour), such as a *Barchan Dune* unit (Exhibit 10), much of the Humanities material was inferential and interpretive.

In comparison with later HASS tests, CSSE Humanities had little of the more formal social science or much tabular or graphical material. In the school curriculum of the 1960s and 70s there was greater emphasis on the humanities aspect, leaving more formal social science to tertiary education.

What is tested in the CSSE Comprehension and Interpretation (Humanities)?

As the name implied, CSSE Humanities was fundamentally a matter of comprehension and interpretation. While there was a good deal of reading, and some of it was very complex, there was also a good deal of interpreting visual and diagrammatic material. Some of the material was contemporary and accessible, but some of it was quite foreign or distant and candidates were expected to deal with a range of unfamiliar allusions. There was emphasis on critical thinking, in terms of comparing different viewpoints and arguments, and on visual and even aesthetic interpretation. The tone and texture of language was often the focus of questions, and there was a good deal of emphasis on thematic and conceptual thinking. Many units prefigured the current emphasis on multimodal representation of ideas and information.

Implicit in the CSSE Humanities was a view that the material used in the test, and even the items, would make a useful basis for thematic inquiry and class discussion. In *From Testing to Teaching in the Humanities,* Audrey Grant and Graeme Withers (1973) discussed the CSSE Humanities of 1971 and showed the richness of the stimulus material, and what it meant to analyse and understand it. They showed the way in which the CSSE units could offer substance for teaching and learning.

How the CSSE Humanities developed over the decade

The early versions of the CSSE Humanities test were concerned to cover as many items as possible, and the test of 1964 had 100 items to be completed in two hours. Candidates would have to work quickly to do so many items in that amount of time. There was also experimentation in the early years to find the most efficient response formats.

Over time there were fewer units and fewer items in the CSSE Humanities. The first CSSE Humanities unit was a brief passage and five items (see Exhibit 1). On the other hand, the creation myths unit of 1973 had three prose passages, a verse passage and four paintings. There were 17 questions about this material. Candidates were to look at issues from multiple perspectives in the

CSSE. Response formats became more consistent but questions became even less formulaic over the decade of testing.

The production values of the test were significantly raised over the life of the test. The foolscap sized test of 1964 had a Dickensian appearance. There was very little white space, and items ran on rather than placing units on separate pages, as in the later versions. Art work was rudimentary in the early tests, but by 1970 a designer had been at work, and the test even had its own logo. While the test was still monochrome, separate colour sheets were included, as for the painting in the creation myths unit.

Criteria for the assessment of Humanities items

This discussion is punctuated with various CSSE items. The following comments aim to describe the Humanities examples and to present a view of what characterises the best cross-curricular Humanities material. Superior cross-curricular Humanities items aim to ask real and authentic questions that are meaningful and significant. The emphasis is on reasoning rather than knowledge, with the aim to elicit different kinds of reasoning. The best cross-curricular Humanities items are open-ended and higher-order because they deal with sub-text, implications, meaning and significance. Such items arise from a diverse range of rich and interesting material, and the questions asked are fundamental and central. The items often require intellectual flexibility, and the ability to take a critical stance. With superior cross-curricular Humanities items, candidates need to demonstrate insight as well as logico-deductive decision making.

Exhibit 38.
Battle of the Somme, 1969
Unit 5, Items 49-58

The following are five accounts of the Battle of the Somme.

Passage 1

Such were the plans for the battle in which the new British Army was for the first time to play the leading part. The conditions were not all entirely favourable. Haig had not been given the full time which he thought desirable either for training the divisions or for accumulating guns and ammunitions.

The general conditions, therefore, under which the British Army was to be launched into its greatest fight, while not entirely favourable, were undoubtedly far more so than those under which previous offensives had been undertaken. As at Loos, the most favourable factor was probably the spirit of the new divisions themselves. Of the great British force practically the whole - whether in old formations or new - belonged to the new British army, raised and trained since the outbreak of the war. The troops themselves were still volunteers, inspired by a pure sense of duty and patriotism, high courage and noble idealism.

Passage 2

What it was like on the day of battle cannot be imagined by those who were not there.

It was a day of intense summer beauty, full of roaring violence and confusion of death, agony and triumph, from dawn till dark. All through that day little rushes of men of our race went towards that No Man's Land from the bloody shelter of our trenches. Some hardly left our trenches, many never crossed the green space, many died in the enemy wire, many had to fall back. Others won across and went further, and drove the enemy from his fort, and then back from line to line and from one hasty trenching to another, till the Battle of the Somme ended in the falling back of the enemy army.

Passage 3

With this brilliant action the Battle of the Somme - the hardest and bloodiest ever fought by the British Army - ended. The British part in it was the logical outcome of dull, determined strategy and the devotion of an inexperienced army. Almost the whole of the 500,000 British troops who - according to calculations made before the battle - were available for expenditure in casualties, were duly expended; and the question arises, how far

that sacrifice was justified by the results. Haig claimed 'the enemy's strength had been very considerably worn down'. But his contention, that the offensive had been intended to wear down rather than break through, is unlikely to be upheld by posterity.

Passage 4

It is inconceivable that Haig would have persisted in his offensive on the Somme had he realised, even approximately, how much lighter than his own was his enemy's loss; nor would the Government have permitted him to do so.

The truth appears to be that, although Haig did believe attrition to be necessary, it was, in this battle, merely a subsidiary aim. From first to last he was big enough to adhere to his intentions of 'breaking through'. He failed to break through and because he failed his literary supporters have argued that it was never his main purpose; if that were true - which it is not - the most comprehensible reason for his conduct of the battle would disappear.

Passage 5

Haig's mind was thoroughly orthodox and conventional. He does not appear to have had any original ideas; no one can discern a spark of that mysterious, visionary, often sinister genius which has enabled the great captains of history to dominate the material factors, save slaughter and confront their foes with novel apparitions. There were the Germans in their trenches. Here he stood at the head of a group of mighty armies. Hurl them on and keep slogging at it in the best possible way - that was war. It was undoubtedly one way of making war, and in the end it was certainly overwhelming victory. But these truisms will not be accepted by history as exhaustive.

49. Which one of the five passages is least concerned with offering an explanation of the Battle of the Somme?

A.1 B. 2 C. 3 D. 4 E. 5

50. These five passages show that some points connected with the Battle of the Somme are debatable. Which one of the following points do the passages interpret differently?

A. Haig's preparation for the Battle.
B. Haig's desire to break through the enemy lines.
C. Haig's relationship with his troops.
D. Haig's desire to win.

51. Passage 3 points to what the author sees as an uncertainty in Haig's aims in this battle. The passage which accounts for Haig's actions in such a way as to remove the uncertainty is

A. Passage I B. Passage 2 C. Passage 4 D. Passage 5

52. The final sentence of Passage 1 states that the troops were 'inspired by a pure sense of duty and patriotism, high courage and noble idealism'. The Passage which gives the best specific example to support this statement is

A. Passage 2 B. Passage 3 C. Passage 4 D. Passage 5

53. The author of Passage 5 implies that Haig's conduct of the battle was, in the main, justified by the result. In the last Sentence, however, he expresses a doubt in the form of a prediction. In which of the following passages is the same doubt most clearly expressed?

A. Passage I B. Passage 2 C. Passage 3 D. Passage 4

54. Passage 5 comments that the Battle of the Somme ended in 'overwhelming victory'. Which of the following statements based on information given in the passages gives most confirmation to this comment?

A. The Battle of the Somme ended with the falling back of the enemy army.

B. Haig claimed that the military strength of the enemy had been considerably worn down by the Battle of the Somme.

C. Throughout the Battle of the Somme, Haig intended to break through the enemy lines.

D. The conditions under which the British were fighting were more favourable than those for any previous engagements.

55. Which one of the following statements from the passages offers an explanation for the British losses which would be the most difficult to support by factual evidence?

A. The army was inexperienced and there had not been all the time necessary for the training and equipping of the troops, and the accumulation of munitions. (Passage 1)

B. It was calculated before the Battle of the Somme that Britain had five hundred thousand troops that she could afford to lose in the battle. (Passage 3)

C. Estimates that Haig had of the losses sustained by the Germans during the Battle of the Somme were inaccurate. (Passage 4)

D. Haig lacked the spark of mysterious genius which had enabled the great generals of history to dominate material factors, and save slaughter. (Passage 5)

The Passage below refers to the British and German loses during the Battle of the Somme.

Passage 6 (Questions 56-58)

It is now known that on November 15th, 1916, the French, Russian and British staff agreed in estimating the German casualties on the Somme at 630,000 against 485,000 for those of the French and British. A very different notion as to the relative loss was held by some careful observers at the front. But it was not until the summarised casualties of the opposing forces were officially published after the war that the extent of the tragic error was known.

Below are the official post-war estimates.

Whole Western Front 1/7/16—19/11/16:
British losses 463,000
Germans 218,000

Estimate for Somme by deduction of losses on other part of the front:
British 410,000
Germans 180,000

56. The phrase 'the extent of the tragic error' in this context has been used by the writer to indicate that

A. German casualties in the Battle of the Somme were more than half the total German casualties on the Western Front.

B. the number of British casualties was overestimated.

C. the British offensive was based on mistaken estimates.

D. lives were wasted through a campaign based on misguided estimates of what was being achieved.

57. Passage 6 implies that

A. it was difficult to estimate casualties because the staffs of different nations were involved.

B. some observers at the Front could have estimated the numbers killed more accurately than the estimates given by the French, Russian and British staff.

C. the French, Russian and British staff had sustained such heavy losses that they were out of touch with the situation at the Front.

D. some observers at the front withheld information on the relative losses.

58. Below, taken from Passages 1 - 5, are several comments on Haig's conduct of the Battle of the Somme. Which one of them has most doubt cast on it by information given in Passage 6?

A. Haig claimed that the enemy's strength had been very considerably worn down. (Passage 3)

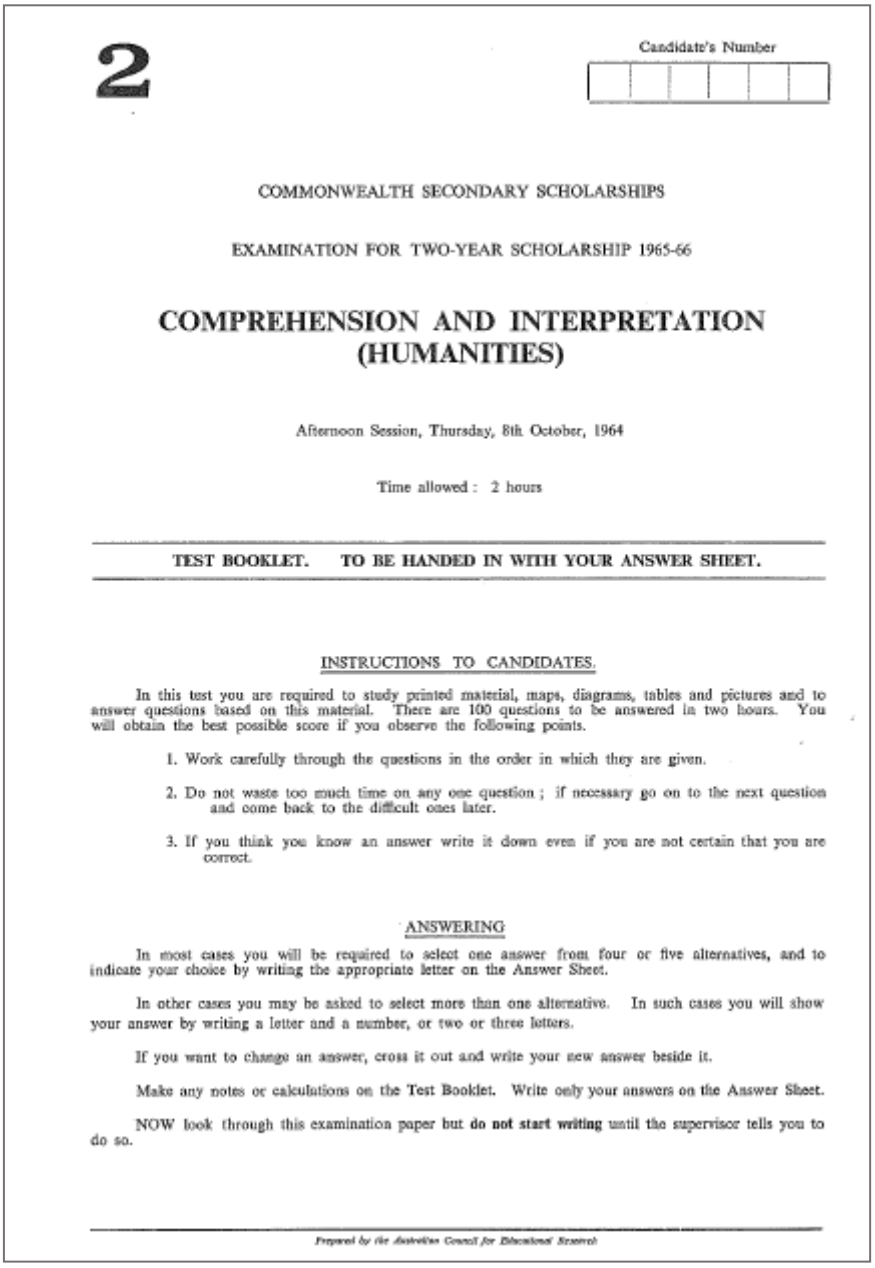
B. Haig intended to 'break through' the enemy line. (Passage 4)

C. Haig hurled his troops at the opposition and kept them slogging at it. (Passage 5)

D. Haig had not been given all the time he would have liked for accumulating munitions and training the troops. (Passage 1)

Some CSSE Humanities units

While there was an evolution in the Humanities material over the 10 versions of the CSSE, the first version of the test was not conservative. This can be seen from the reading units of the first Humanities test of 1964, presented on page 16. The front cover of the 1964 Humanities test is presented in Figure 7.



2

Candidate's Number

COMMONWEALTH SECONDARY SCHOLARSHIPS

EXAMINATION FOR TWO-YEAR SCHOLARSHIP 1965-66

COMPREHENSION AND INTERPRETATION (HUMANITIES)

Afternoon Session, Thursday, 8th October, 1964

Time allowed : 2 hours

TEST BOOKLET. TO BE HANDED IN WITH YOUR ANSWER SHEET.

INSTRUCTIONS TO CANDIDATES.

In this test you are required to study printed material, maps, diagrams, tables and pictures and to answer questions based on this material. There are 100 questions to be answered in two hours. You will obtain the best possible score if you observe the following points.

1. Work carefully through the questions in the order in which they are given.
2. Do not waste too much time on any one question ; if necessary go on to the next question and come back to the difficult ones later.
3. If you think you know an answer write it down even if you are not certain that you are correct.

ANSWERING

In most cases you will be required to select one answer from four or five alternatives, and to indicate your choice by writing the appropriate letter on the Answer Sheet.

In other cases you may be asked to select more than one alternative. In such cases you will show your answer by writing a letter and a number, or two or three letters.

If you want to change an answer, cross it out and write your new answer beside it.

Make any notes or calculations on the Test Booklet. Write only your answers on the Answer Sheet.

NOW look through this examination paper but **do not start writing** until the supervisor tells you to do so.

Prepared by the Australian Council for Educational Research

Figure 7. Front cover of the first CSSE Humanities test of 1964

The first Humanities unit in the CSSE was a brief 140 word prose passage we will call *The Drone of the Tambura* (Unit 1, 1964, page 16). This unit tests reading comprehension and interpretation, but the items are not literal comprehension or information retrieval. The first item asks for a global summary of the writers 'sensations'. The second item is about the implications of the writer's statements. The third item asks an adventurous question about what is 'chiefly notable' about the passage. (Contrast this item with the 'evaluation' item from the SAT on page 56.) A literal minded

reader might not know what it means to say that something in a text is more or less 'notable' than other things. Such literal mindedness is at a marked disadvantage in the CSSE Humanities. The fourth question asks about the 'style of the passage', and the answer to this question is neither simple nor clear.

The sixth item asks candidates to consider four rising and falling lines as a means of depicting the 'pace' of the passage. As with many similar CSSE items, what is meant by 'pace' in the item is neither clear nor simple. The candidate has to construct a possible meaning for the word 'pace' in the question.

As a test of reading skills *The Drone of the Tambura* is not information retrieval or literal comprehension. It deals with tonal and textural issues at a high level of abstraction. Most of the items are adventurous challenges to think.

This first unit in CSSE Humanities can be used as a benchmark for sophisticated reading comprehension. This small first unit is prefiguring the kind of reading and thinking to be seen in the very diverse sets of material used in later tests.

The last unit in the Humanities test of 1964 is on a larger scale than the first, but it is still comparatively modest in comparison with later tests. The *Historians on the Norman Conquest* (Exhibit 6, page 18) presents the contrasting views of Professors Sir Chas. Oman and H. W. C. Davis. (Attention is drawn explicitly to their academic titles so as to focus the issue of conflict in authoritative comment.) Both texts are very complicated in that they are constantly referring to unknown figures and arcane matters. Candidates have to be able to hold their nerve and keep reading when they don't know who someone or what something is. This ability to infer and hold meaning in spite of unknown references is fundamental to higher level reading. No attempt is made to mitigate the difficulty of such allusive texts in CSSE. There is very little footnoting of difficult words or unfamiliar ideas or events in CSSE Humanities.

Item 88 asks the central question about the purpose of passage 1 of the *Historians* unit. Item 89 is about the significance of the phrase "splendid twilight" in passage 2, and that is an indirect way of asking a question about the purpose and substance of that passage also. Items 90-93 ask candidates to use a key that categorises statements as fact or opinion, and as likely to be made by or as supporting the view of either writer. Item 94 is a complex setup asking candidates to draw an overall conclusion and nominate a reason for the conclusion. The strictest answer to that question would be that: "They are equally likely to be right" although the justification to match this with "they are both professors of history" is unsatisfying. Failing this strict conclusion, the best response would seem to be Passage I because "he gives more evidence to support his views". Again this adventurous conclusion is not satisfying because a 394 word passage has more opportunity to make factual claims than a 207 word passage. This is very adventurous multiple-choice question testing, and it is arguable that this item goes over the edge of plausible reasoning into speculation. Items 95-100 are on safer ground in asking if a number of statements are likely to gain assent or provoke dissent from one or other professor.

This unit is very difficult reading and interpretation, the response formats are complicated, and some of the answers are arguable. They are very adventurous multiple-choice items.

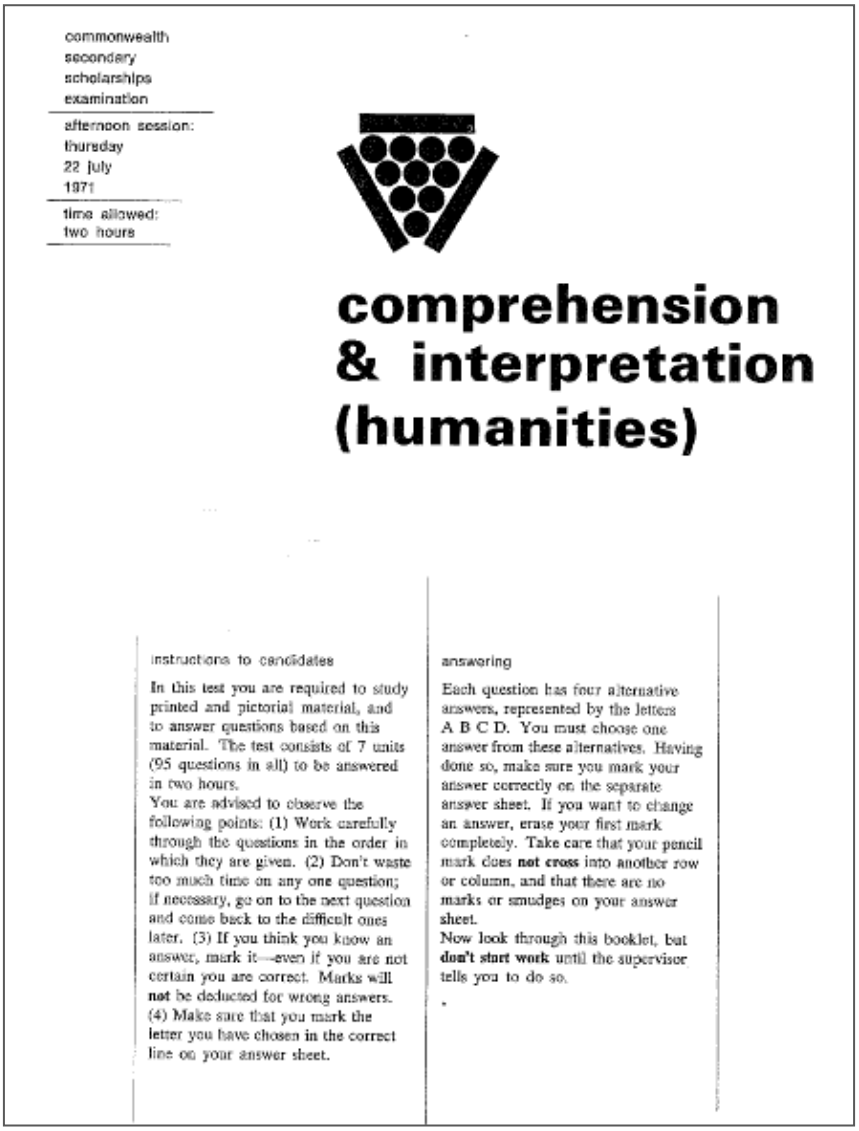
CSSE Humanities began with challenging items. In some ways the Humanities items became a little tighter or more conservative over time. The response formats also became more consistent and simpler, and there was less challenge for candidates in negotiating the response format itself. The later versions of the test are cleaner and tighter, but they are still very adventurous.

While there was a significant emphasis on inferencing and interpretation in the Humanities material, there was also some information processing and literal comprehension. Exhibit 10, *The*

Barchan Dune unit in the first test (Unit 12, 81-87, 1964, page 51) involved understanding a written text about a curious kind of sand dune, and matching that information with a pair of diagrams. What one has to do in the unit is quite clear and transparent, but it is genuinely difficult to do. The way a Barchan Dune is structured and moves as described in the text feels counter-intuitive and mysterious.

The Barchan Dune unit is not all literal transposition. As well as working out from the text and diagrams how the dune is formed and moves, candidates are asked to identify the 'chief impression' given by the writer. Item 83 asks candidates to construe the phrase "in their blind advance", and such qualitative interpretation is extended in Item 85 into typifying the 'approach' of the writer. While the unit is fundamentally literal comprehension, there are also a number of impressionistic questions.

The front cover of the 1971 Humanities test is shown in Figure 8. From this test, Exhibit 39, the *Marshall McLuhan* unit (Unit 7, 1971, page 121) presents some very complex comments by the celebrated theorist about media and education. These ideas were some of the most advanced of the time, and they have survived very well. Candidates were asked to construe terms in the text, and see the implications of various statements.



commonwealth
secondary
scholarships
examination

afternoon session:
thursday
22 july
1971

time allowed:
two hours

comprehension & interpretation (humanities)

instructions to candidates

In this test you are required to study printed and pictorial material, and to answer questions based on this material. The test consists of 7 units (95 questions in all) to be answered in two hours.
You are advised to observe the following points: (1) Work carefully through the questions in the order in which they are given. (2) Don't waste too much time on any one question; if necessary, go on to the next question and come back to the difficult ones later. (3) If you think you know an answer, mark it—even if you are not certain you are correct. Marks will **not** be deducted for wrong answers. (4) Make sure that you mark the letter you have chosen in the correct line on your answer sheet.

answering

Each question has four alternative answers, represented by the letters A B C D. You must choose one answer from these alternatives. Having done so, make sure you mark your answer correctly on the separate answer sheet. If you want to change an answer, erase your first mark completely. Take care that your pencil mark does **not cross** into another row or column, and that there are no marks or smudges on your answer sheet.
Now look through this booklet, but **don't start work** until the supervisor tells you to do so.

Figure 8. Front cover of the CSSE Humanities test of 1971

Item 90 focuses on the way language is used in the passage, and candidates had to identify a quotation that is appropriately described as "flamboyantly indecipherable". Rather than being asked what a statement means, this unusual and elegant item asks candidates to identify a statement of uncertain meaning.

Item 91 offers four diagrams and asks which best represents the "way ideas are put together within the passage". At first the significance of this question is hard to see, and the answer is quite obscure. The central focus of Diagram A and the linear focus of Diagram D can be eliminated from contention. Diagram B seems less complicated than the passage, whereas Diagram C suggests the overlapping complexity of the passage.

A New Yorker cartoon by Alan Dunn about Marshall McLuhan is then introduced, and the questions require candidates to relate the cartoon to the passage and interpret what is shown in the cartoon and suggested by the complicated caption about the attitudes of the father and son. These items on the cartoon are central to the issues in the unit and involve sophisticated interpretation of both the text and the cartoon by candidates who had probably never heard of McLuhan or seen The New Yorker.

There are pedestrian units and items in the CSSE, and in a sense this is inevitable, but there are many sophisticated, subtle and adventurous units and items in the CSSE Humanities.

Exhibit 39.
McLuhanism, Humanities 1971
Unit 7, Items 81-95

This unit consists of a passage by Marshall McLuhan and a cartoon from the New Yorker.

There is a world of difference between the modern home environment of integrated electric information and the classroom. Today's television child is attuned to up-to-the-minute 'adult' news - inflation, rioting, war, taxes, crime, bathing beauties - and is bewildered when he enters the nineteenth-century environment that still characterizes the educational establishment where information is scarce but ordered and structured by fragmented, classified patterns, subjects, and schedules. It is naturally an environment much like any factory set-up with its inventories and assembly lines.

The 'child' was an invention of the seventeenth-century; he did not exist in, say, Shakespeare's day. He had, up until that time, been merged in the adult world and there was nothing that could be called childhood in our sense.

Today's child is growing up absurd, because he lives in two worlds, and neither of them inclines him to grow up. Growing up that is our new work, and it is total. Mere instruction will not suffice.

The young today live mythically and in depth. But they encounter instruction in situations organized by means of classified information - subjects are unrelated, they are visually conceived in terms of a blueprint. Many of our institutions suppress all the natural direct experience of youth, who respond with untaught delight to the poetry and the beauty of the new technological environment, the environment of popular culture. It could be their door to all past achievement if studied as an active (and not necessarily benign) force.

The classroom is now in a vital struggle for survival with the immensely persuasive 'outside' world created by new informational media. Education must shift from instruction, from imposing of stencils, to discovery — to probing and exploration and to the recognition of the language of forms.

The young today reject goals. They want roles - R-O-L-E-S. That is, total involvement.

They do not want fragmented, specialized goals or jobs.

Marshall McLuhan: from *The Medium is the Massage*

81. This passage indicates that the present school environment is unsuitable for growth. The reason implied for this in the passage is that schools

A. refuse to structure learning in a sufficiently organized manner.
B. do not present the student with properly classified information.
C. fail to provide the student with opportunities for discovery and creative interaction.*
D. do not present the student with a sufficient quantity of information.

82. Which one of the following is most essential to 'growing up' as it is presented in paragraph 3?

A. Becoming wise.
B. Being involved.*
C. Rejecting instruction.
D. Absorbing information.

83. At several points in the passage institutionalized education is criticized. Which one of the following is not given as a criticism of institutionalized education?

A. It organizes and structures.
B. It is concerned with probing and exploring rather than giving information.*
C. It rejects total involvement.
B. It deals with classified patterns and subjects.

84. Which one of the following views of the nature of motivation in learning is implied by this passage?

A. Pressure produces product.
B. A firm hand produces a disciplined mind.
C. Lead on to the light.
D. This organism is self-activating. Don't restrain it.*

85. In paragraphs 4 and 5 McLuhan indicates what he thinks schools should be doing. Which one of the following phrases is most consistent with what he suggests? School activities should involve

A. tuning in to the environment around them.*
B. inquiring into the nature of life.
C. instruction in the vast store of knowledge that comprises our culture.
D. discovery of the world of myths and heroes.

86. Paragraph 5 refers to the persuasive force of the medium through which information is communicated as compared with the actual content of the information. Which one of the following most accurately indicates the relative emphasis of the medium and the content as it is given here? As a persuasive force the medium

A. plays a greater role than the content.*
B. is just as effective as the content.
C. is less important than the content.
D. is as ineffective as the content.

87. According to paragraph 6, the young today are in search of

A. ideals by which to live.
B. absorbing careers.
C. masks behind which to hide themselves.
D. an identity which they can assume.*

88. According to McLuhan the reaction of 'today's television child' to institutionalized education is that it

A. is bewildering since it forces him into contact with the adult world before he is ready for it.
B. does not engage his interest since it does not use television and other audio-visual means of presenting information.*
C. is incomprehensible since it does not represent the kind of environment to which he is adjusted.
D. is inadequate since it stresses an understanding of the natural environment but not of the achievements of the past.

89. Below are four statements made by McLuhan about school education as it is today. Which one of them most closely matches McLuhan's view of present-day education as it is presented in this passage?

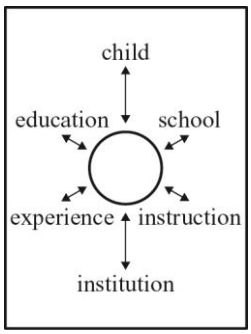
A. 'our current schools are intellectual penal institutions'
B. 'in today's world, the least education is the best education'
C. 'in our schools there is simply too much to learn'
D. 'to expect a "turned on" child of the electric age to respond to the old education modes is rather like expecting an eagle to swim'*

90. Some of McLuhan's statements such as 'The electric light is pure information' have been called 'flamboyantly indecipherable'. Which one of the following quotations from the passage best illustrates this quality?

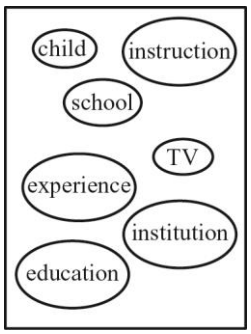
A. 'Today's television child is attuned to up-to-the-minute "adult" news'
B. 'Today's child is growing up absurd'*
C. 'Many of our institutions suppress all the natural direct experience of youth'
D. 'Education must shift from instruction'

91. Of the following diagrams which best suggests the way that ideas are put together within the passage?

A

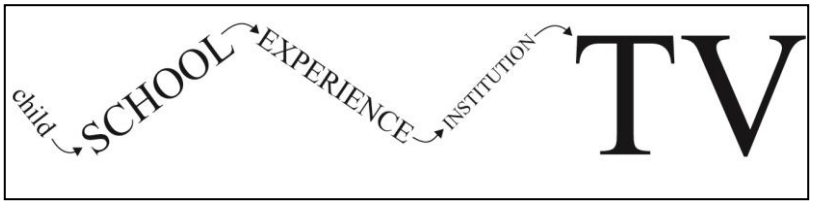


B



C *

D



94. Which one of the following extracts from the passage is best illustrated by the son in this cartoon?

A. '..youth, who respond to the poetry and the beauty of the new technological environment'. **
B. 'Today's television child is attuned to up-to-the-minute "adult" news'.
C. 'The young today encounter instruction in situations organized by means of classified information'.
D. 'Education must shift from instruction, from imposing stencils, to discovery'.

95. Which one of the following extracts from the passage indicates an environment most similar to that represented by the father in the cartoon?

A. 'the modern home environment of integrated electric information'.
B. 'the nineteenth-century environment that still characterizes the educational establishment'.**
C. 'the beauty of the new technological environment, the environment of popular culture'.
D. 'the immensely persuasive "outside" world'.

10

The art of cross-curricular item writing: CSSE Science and Quantitative Thinking

The objectives of the CSSE (1967) make particular reference to Bloom's Taxonomy of Educational Objectives (1956) and the International Association for Educational Assessment mathematics project (1964). Both of these influences encouraged an emphasis on reasoning and problem solving in the CSSE. Their influence can be seen particularly in CSSE Science and Quantitative Thinking. As with the Humanities, CSSE Science and Quantitative Thinking are the beginning of a tradition of testing the generic skills that underpin the study of science and mathematics at ACER, and the approach is still evolving. CSSE is the progenitor of PISA scientific and mathematical literacies and PIAAC adult numeracy.

CSSE Comprehension and Interpretation (Sciences)

The objectives outlined in 1967 focused on science as understanding the material world or the environment.

> *Science is concerned with the relationships between aspects of the environment. Knowledge of such relationships enables us to describe accurately the outcome of interactions within the environment, and thus to some extent to predict, modify or control the outcomes of such interactions.*

The science paper was described in terms of the problem solving skills of:

- classification;
- translation;
- extrapolation;
- application; and
- synthesis.

These skills echo those identified by Bloom's taxonomy.

In retrospect Peter Thomson who was responsible for the Science paper in 1974 described the criteria for including material in the test as determined by subject matter, scientific skills, and interest and relevance. The aim was to present a balanced approach spanning as many scientific subject areas as possible. The material could deal with any field of science and technology, or it may be related to a scientific investigation of a non-scientific field.

According to Thomson, the following scientific skills were tested in CSSE Science.

a) An understanding of the way scientists go about their work (this is sometimes referred to as the 'scientific method').

b) The ability to handle scientific data in all its forms, e.g. reading tables, graphs, flow charts; following circuit diagrams; analysing scientific writing.

c) The recognition of scientific and mathematical models.

d) The classification of scientific information.

The second of these statements is tested in all cross-curricular MST material, but the first, third and fourth characteristics are given less emphasis in later MST testing.

Thomson made the following comment on the importance of 'interest and relevance' to the development of the science paper.

> *It is considered important that the paper does not become one which only selects those candidates who are best able to wade through an accumulation of dull, uninteresting material. Considerable thought is therefore given to providing a paper which can hold a student's interest. We would like to think that some students even enjoy the experience. In the 1972 paper the introductory unit could have equally well been presented in written form, but the 'comic book' approach was felt to have much more appeal.*
>
> *Relevance goes hand in hand with interest. Apart from subject area relevance, an attempt is made to include material closely related to the problems of today's society.*

Quantitative Thinking and Comprehension and Interpretation (Mathematics)

The following objectives were identified in 1965 for the Quantitative Thinking sub-test.

a) Ability to compute and manipulate numbers with skill and accuracy.

b) Knowledge of definitions, notations, operations, etc.

c) Ability to translate between verbal, symbolic, tabular, graphical, diagrammatic and pictorial material.

d) Ability to read and comprehend new mathematical material.

e) Ability to apply known concepts to new problems.

f) Ability to analyse problems and to determine the relationships between their various parts.

g) Ability to follow and construct proofs.

h) Inventiveness - ability to reason creatively and to arrive at tentative, intuitive solutions.

The following comments were made on these objectives that described quantitative thinking as conceptualisation and problem solving. The mathematics knowledge required by the test was modest and elementary, and reasonable approximations and estimations were required for most questions rather than exact calculation.

> *The skills of arithmetical computations and arithmetic and algebraic manipulations are required to varying degrees throughout the study of mathematics. Likewise many of the solutions to the problems in the Quantitative Thinking examinations can only be reached after some computations have been performed. However, most of the figures are contrived so that, while knowledge of the computing process is important, the major difficulty of the question will not be due to the difficulty of the computation. There are one or two questions in which the main objective is to test the ability to compute accurately but in general it is considered*

> *that the important thing is to recognise the order of magnitude of the answer. This category also includes the ability to make reasonable approximations and to apply reasonable computational checks to a solution in order to determine its general appropriateness.*
>
> *Because of the varied backgrounds of candidates, severe limits are imposed on the specific content knowledge that it is reasonable to assume as common to all candidates. No special knowledge of mathematics is required beyond the syllabus for the second year of secondary education in any state. In general, questions testing only knowledge have not been included, but of course some elementary mathematical knowledge is required in all questions. In some cases knowledge of special techniques or generalisations, etc., may lead to a more elegant solution but, in general, this possibility is avoided.*
>
> *The production of, prediction from and verification of suitable scientific models of the environment presuppose certain problem solving skills (amongst other skills and abilities) in the worker in the field. These problem solving skills should be present, although probably in a less developed form, in the student of science, and in still less developed form in an educated 'layman' interested in science.*

In 1973, David Arney, who was then responsible for the paper, described three basic criteria used in selecting material for what had then changed from the Quantitative Thinking paper to the Comprehension and Interpretation (Mathematics) paper: subject area, mathematical skills, and interest and elegance. The stimulus material was drawn from as diverse areas of mathematics as possible, including applying mathematics to 'scientific' or 'practical' data. It could include arithmetic, number theory, graphs of various types, algebra, plane geometry, group theory, spatial relations and 3D geometry.

According to Arney, the specific skills required to answer the questions were:

a) computational ability;

b) the ability to deal with all forms of mathematical data - numerical, tabular, diagrammatic, graphical and verbal;

c) the ability to apply given or previously learned information to find the solution of problems; and

d) the ability to analyse problems and see relationships between parts of the data.

In terms of 'interest and elegance', Arney considered it important that the material be as interesting as possible because "many students are not fond of mathematics". He ruefully added that the item writers "might only manage to interest a minority", and that the "classification of an item or unit as interesting or elegant is usually a very subjective judgment, although there is considerable agreement on boring questions".

While there were some discrete items in the early versions of the Science and Quantitative ability, the tests were based on stimulus material or topics and a set of diverse questions about that material. The aim was to eschew classroom science and mathematics and the abstract skill items of the psychometric approach. Candidates were challenged with unfamiliar material, and they were commonly expected to construct answers to unique questions using a non-standard process.

The stimulus tended to be topical and contemporary, and candidates had to think about issues and make sense of them. The stimulus was not a just a 'context' for the items, it was the substance of the thinking expected of candidates.

Just as the stimulus material was very diverse, so the kinds of questions asked and the thinking required were also very diverse. Candidates had to demonstrate intellectual flexibility in understanding a topic or activity. The questions seemed to be conceptual (about understanding the material or the task) rather than procedural (working out the right answer to a question). Questions

seemed to require global understanding of the material rather than merely working out the answer to this or that question. The approach was teacherly, and at times it seemed to be a kind of teaching through exploration of the material. There was also an interest in and promotion of science and mathematics as socio-cultural activities.

In retrospect the CSSE Science and Quantitative ability material are very challenging. Candidates were made to work to understand the material, it was not made easy for them, and at times they even seemed to have to discover something rather than merely work out the correct answer.

In CSSE, science seems an activity in which questions are raised and answers considered. This science is a process of inquiry and discovery, and science is not only a body of settled knowledge. Science is theory construction in CSSE, and there is uncertainty and dispute in the scientific thinking considered in much of the material.

Units could require candidates to consider the assumptions and reasoning that would lead to a certain conclusion, and they might be required to consider alternative interpretations or conclusions. There was also an interest in how conclusions are drawn and questions answered in CSSE science. Some CSSE units were quite exploratory and adventurous, and in retrospect some of the questions might be seen as too adventurous.

There is a clear tendency in CSSE Science towards inferential thinking and fuzzy reasoning. There are quite a few answers to CSSE questions that are not calculations or matters of logical necessity. The questions can be hypothetical (even speculative), and the answers are not always irrefragable. Some CSSE questions can only be answered in terms of plausibility, some options offered are very close together, and there is room for argument about the 'best answer' in quite a few Science items.

Just as the emphasis in CSSE Science was on thinking conceptually about the stimulus, so the Science material expected high levels of comprehension and interpretation of candidates. Units could spread over a number of pages, and have a good deal of written text with a number of interrelated questions. It can seem in CSSE Science and Quantitative ability that if the candidate can't grasp the whole, they will have difficulty with all the items. Individual items are not presumed to be discrete or independent. Often in Science and Quantitative ability units it can be necessary to work from the stimulus towards the items, where as in some more modern MST material candidates are wise to work from the items back to the material, or the relevant part of the material. It can sometimes seem that they don't have to grasp the whole to answer the questions.

While CSSE Quantitative ability papers contained some recognisable mathematics, in many cases the questions involved logical thinking rather than the abstract mathematical approach usually seen in standard mathematics test papers. The stimulus material was often based on real contexts and there were the occasional 'proof' or 'theorem' style stimuli which involved logical reasoning.

The diverse stimulus used for the Quantitative ability units included network diagrams, tables of voting data, sales based information, medal tallies that involved logico-quantitative thinking and some abstract mathematics. The clear focus was on items that were not formulaic or algorithmic, and which would draw fully upon the stimulus material. In both the Quantitative ability and the Science, more so in the Science, the stimulus material brought current ideas (for the time) to the students. It was not uncommon to see contemporary real-world material on the effect of a particular drug, a psychological study, underwater communication cable breaks, or applications of sales and marketing techniques. While the Quantitative ability items appear less adventurous than the Science items, they were also focused on reasoning skills and conceptual understanding rather than knowledge.

At its best the CSSE Science and Quantitative ability is real world, concrete and meaningful. Real scientific issues and actual data are valued as stimulus material. At its best real and authentic scientific and mathematical thinking is elicited in CSSE Science and Quantitative ability, and

candidates have to think like mathematicians or scientists. The material aims to be interesting, and is at least not routine. The items require intellectual flexibility, and it was as though candidates have to learn how to do them.

There is a concern with science as an activity and the scientific method and theory construction in CSSE Science. And there is a concern with mathematics as a way of knowing to be seen in CSSE Quantitative ability.

The best CSSE Science and Quantitative ability is both concrete and conceptual. It is concrete in that it is about something, and it is conceptual in that it is about understanding the topic as a whole. There is a clear interest in hypothesising, inferencing and extrapolating in CSSE Science and Quantitative ability. The best CSSE Science and Quantitative ability questions are:

- centrally focused;
- require global understanding of the material;
- are related to the interest and value of the stimulus;
- are conceptual rather than information processing;
- have substantive rather than nugatory difficulty; and
- place more emphasis on conceptual understanding than procedural reasoning or calculation.

Some CSSE Science and Quantitative thinking units

A selection of Science and Quantitative ability units, in whole or in part, are presented in each chapter of this text as indicative of the CSSE approach. Some comments on this material are offered below. Some of these comments are a critical dialogue with the writers of the past. Even when they don't know it, the item writers of the present are interacting with the past. At its best the art of item writing is a matter of emulation and competition with a tradition.

Exhibit 2. The Aspirin Age, Science 1972 (see p.7)

While one would have to conclude that the topic and the presentation of it are exemplary, one would note in the due spirit of emulation that the questions do not do justice to the material. The questions are not much more than literal comprehension. The possible relationship of prostaglandins to contraception, for instance, is mentioned but not explored. There should be a question about the suggestions and significance of the third and second last panels of the strip. Those comments and images seem provocative from a contemporary perspective. As is often the case, these items may have been written and trialled, but they might not have attained statistics of a satisfactory difficulty or discrimination.

Exhibit 3. The Centre for Short-lived Phenomena, Science 1970 (see p.8)

The genesis of this unit is discussed on page 8. As well as being first hand and contemporary science, the unit shows the problems and pitfall of scientific observation. Two questions ask how observations lead to conclusions. Two questions are about potential differences of interpretation. The last question is a kind of socio-cultural speculation.

Exhibit 7. The 'Waggle Dance', Science 1969 (see p.28)

This unit is based on the curious way bees return to the hive and relay information on the location of food to the population of the hive via the 'waggle dance'. The behaviour of the bees is complicated and candidates are asked to match descriptions of the activity with diagrams, and the reverse. One has to understand the behaviour and understand how it might be represented. The candidate has to imagine the behaviour of the creature in space, and the information is genuinely difficult to understand.

Exhibit 8. The Olympic medal tally, Quantitative thinking 1970 (see p.30)

This Quantitative unit is about the different ways an overall result or ranking may be calculated from different tiers of medals. The issue is evident, accessible, concrete, and significant. There is an intrinsic and (in 1970) an immediate interest in the overall results of the Olympics of 1968. Candidates have to make numerical sense of concrete changes in relativities. The candidates have to find a way of manipulating the data to efficiently answer the questions.

This is an example of the different ways numbers can be used. The calculations are simple but how the calculations work is quite complicated. The numbers have meaning in terms of the overall conception of winners and losers. Compare the multiple-choice version of 1970 with the short response version of similar ideas in 2015. The short response version can get much closer to the material than the multiple-choice version.

Exhibit 15. The date of Easter, Quantitative ability 1968 (see p.60)

In this unit candidates are given the rules for determining the date of Easter Day, and asked to use them in various conditions. It is a real and complicated calculation. Difficult as it is, the questions remain at the level of working out the first and last possible days for Easter. There isn't a question that asked (or one that 'worked'?) about the nature and cause of such variation.

The election count grid is concrete data that candidates have to understand as a whole and apply the structural rules for determining winners. It has the missing data trick (what data would be in the blank cells?), and there are no questions about the validity of the method as such. Again such items may not have 'worked statistically'.

Exhibit 16. Election count, Quantitative ability 1966 (see p.60)

The election count grid is concrete data that candidates have to understand as a whole and apply the structural rules for determining winners. It has the missing data trick (what data would be in the blank cells?), and there are no questions about the validity of the method as such. Again such items may not have 'worked statistically'.

Exhibit 17. Birth rates graph, Quantitative ability 1970 (see p.61)

This complicated presentation of data relates birth and death rates. The presentation is possible because the birth rate is consistently higher than the death rate. The first item is a straight reading exercise. The second question involves a comparison of the different variables to draw a conclusion about population increase. The last question is a matter of plausible inference rather than a calculation.

Exhibit 18. Wheat production graphs, Quantitative ability 1964 (see p.62)

This unit is a complex overlapping pair of graphs from which candidates have to make inferences. The first question is a calculation. The second two questions are conceptual in that they involve a recognition that when the lines diverge they show something about the number of acres sown at that time. These questions seem unintelligible until one sees the point.

Exhibit 19. Types of shell on a beach figure, Quantitative ability 1966 (see p.63)

The graph about the types of shells is an unusual way of presenting information. The unit has substantive meaning and candidates draw conclusions from it, but the items do not get to an overall interpretation of the material.

Exhibit 20. The measurement of population size 1971 (see p.69)

The unit presents different methods of making complex estimations in biological research, and asks candidates to understand and use them through two investigation scenarios. Candidates have

to deal with complex technical concepts, use a classification system, recognise issues for argument and debate, and consider the reliability of different approaches.

The unit feels like real and difficult thinking. The questions are concerned with reasoning about possible conclusions. Candidates also have to interpret algebraic and graphical representations of the methods.

Exhibit 21. Topographical transformations 1971b (see p.71)

Some of the best CSSE Quantitative thinking items involve a conscious and conceptual kind of spatial reasoning, as in Items 53 and 54. These examples involve complex spatial imagination, but unlike spatial reasoning in the psychometric tradition it has substantive meaning rather than being a test artefact or a 'laboratory task'. Exhibit 21 introduces the notion of topographical equivalence and ask candidates to apply it visually and through language.

Exhibit 22. Paper folding, Science 1971 (see p.71)

This unit deals with the genuine complexity of how larger sheets of paper are printed and folded into a series of pages. It is a real and complex problem in the paging of printed documents, and poses genuine challenges to logic and spatial imagination.

Exhibit 23. Burning in air, Science 1964 (see p.72)

Exhibits 19 and 20 are a kind of hypothetical deductive thinking that is quite common in CSSE science. In Exhibit 23, candidates are offered two theories to explain the burning of material in air. The candidates are given a set of facts and asked to decide if a fact supports one or other or both theories, or if it offers support to neither. The unit is not concerned with the correct interpretation, it is concerned with the process of interpretation.

Exhibit 24. High temperature, Science 1967 (see p.73)

This unit involves a kind of hypothesising and theory testing about the relationship between illness and high temperature. Candidates have to assess the plausibility of suggestions as explanations of high temperature. The unit is concerned with inference and theory construction. It is a concrete kind of thinking involving plausible inferences rather than matters of fact

Exhibit 40. Seven segment display, Quantitative 1973 (see p.131)

The 'figure-of-8' display of digital numbers was comparatively unfamiliar in the 1970s. Candidates are introduced to the "0 off, 1 on" convention and the associated seven digit code for a particular display, which they then had to manipulate. Again this was concrete and meaningful exercise in logic and numerical processing.

Exhibit 41. Hypothesis testing, Science 1972 (see p.131)

This unit has an introduction explicitly dealing with the hypothetical deductive method in science. The unit is about hypothesis testing and experimentation. Candidates are to see how different experiments test hypotheses, and what can be concluded from different outcomes of the experiments. The material is concrete and complex, and candidates have to think like scientists.

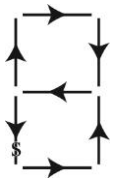
Exhibit 42. Plant evolution, Science 1967 (see p.133)

This and the next unit are conceptual and inferential. In this unit candidates have to find a relationship between some characteristics of evolutionary development and the description of five plant groups so as to infer the sequence of development. Candidates have to interpret the criteria, and make comparisons to hypothesise, draw inferences and make interpretations.

Exhibit 40.
Seven segment display, Quantitative 1973
Items 24-26

A certain computer has seven elements some or all of which light up. They light up in combinations to give the patterns shown below, for the numerals 0-9.

The computer 'reads' the patterns in the order shown, by the arrows in the diagram below. If an element is off (unlit) it is read as 0 and if an element is on (lit up) it is read as 1; so, for example, the number 4 would be read 0011101.



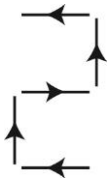
24. How would the number 2 be read?

A. 1010101 C. 1111100

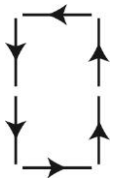
B. 1101011 B. 1101101

25. If the computer's reading sequence breaks down and not all the elements are read (and those which are read may be read in an order which is different from the usual order), it may still be possible to decide from the incomplete display which numeral was intended. Which one of the following reading patterns would allow this?

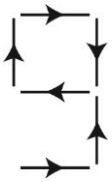
A.



B.



C.



D.

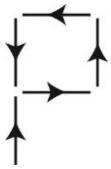


Exhibit 41.
Hypothesis testing, Science 1972

The advance of scientific knowledge is based on experimentation. Observations are used by scientists as the basis for making hypotheses. A hypothesis is a supposition which is put forward to explain certain observations. Any hypothesis put forward should be consistent with all available observations.

In order to test a hypothesis, it must be used to predict observations (results) in new experiments.

If the observations in any new experiment are not consistent with the results predicted using the hypothesis, then the hypothesis is rejected.

If the observations of a new experiment are consistent with the results predicted by using the hypothesis, this does not prove the hypothesis to be correct, as it is possible that some other new experiment may prove the hypothesis to be inadequate; rather this lends support to the hypothesis.

Any supposition which cannot be tested by this method of predicting observations in new experiments is not a scientific hypothesis.

Today a great deal of scientific work is carried out in circumstances where there is only a limited number of 'possible' hypotheses to explain a particular observation. In cases such as these, when one of the competing hypotheses is rejected, the original hypothesis is said to be supported. You are to assume that this is true of the hypothesis described in the passage below.

Under natural conditions, the major species which preys upon the freshwater mosquito fish (*Gambusia affinis*) is a larger fish, the pickerel (*Esox americanus*). If Esox is introduced into an aquarium containing a school of Gambusia, it is observed that Gambusia react in a characteristic way, swimming ungrouped at the surface with fins erect.

A biologist studying the relationship between the two species formed the hypothesis:

'Gambusia react to one or more chemical substances which pass into the water from the body of Esox'.

Among the experiments he performed when investigating this hypothesis was the following:

Experiment 1: Esox was placed into an aquarium for one hour; then, after removing Esox from the water, a group of Gambusia was introduced into the aquarium.

34. When Experiment 1 was performed, it was found that when Gambusia were placed in the aquarium from which Esox had recently been removed, the fish swam ungrouped at the surface with fins erect. This observation

A. supports the hypothesis, but does not prove it to be correct.
B. proves the hypothesis to be correct.
C. proves the hypothesis to be incorrect.
D. does not provide data which is relevant to this particular hypothesis.

35. Which of the following beat describes the relationship between Experiment 1 and the hypothesis?

A. The experiment is designed to provide data on which to base a new hypothesis.
B. The experiment is designed to check on the accuracy of the original observation on which the hypothesis was based.
C. The experiment is designed to test a prediction made from the hypothesis.
D. The results of the experiment will either prove or disprove the hypothesis.

After completing Experiment 1 the biologist carried out another experiment.

Experiment 2: Esox was placed into an aquarium for one hour. Alter removing Esox, the water in the aquarium was quickly filtered through activated charcoal (which is known to be capable of absorbing many chemical substances). Gambusia were then introduced into the aquarium.

36. If Gambusia in Experiment 2 behaved normally (i.e. did not swim ungrouped at the surface with fins erect), this observation would

A. provide additional support for the hypothesis, but not prove it to be correct.
B. prove conclusively that the hypothesis is correct.
C. prove conclusively that the hypothesis is incorrect.
D. contradict the results from Experiment 1.

37. If in Experiment 2 the Gambusia had been observed to swim ungrouped at the surface with fins erect, this observation would

A. provide additional support for the original hypothesis but not prove it to be correct.
B. prove conclusively that the hypothesis is correct.
C. prove conclusively that the hypothesis is incorrect.
D. neither support nor disprove the original hypothesis.

Two further experiments were conducted.

Experiment 3. A wooden model of Esox was introduced into an aquarium containing Gambusia.

Experiment 4: Esox in one glass aquarium were placed alongside another glass aquarium containing Gambusia so that the Esox would be seen by the Gambusia.

38. If in both Experiment 3 and Experiment 4 the Gambusia behaved normally (i.e. did not swim ungrouped at the surface with fins erect), this observation would

A. provide support for the original hypothesis.
B. prove that Gambusia reacted to the presence of chemicals which pass into the water from the body of Esox.
C. prove that the original hypothesis is incorrect.
D. neither support nor disprove the original hypothesis.

39. If in all four experiments, 1, 2, 3, and 4, the Gambusia swam ungrouped at the surface with fins erect, which one of the following statements would be justified?

A. The Gambusia react in this characteristic way only if they can see Esox.
B. The Gambusia react in this characteristic way only if chemical substances released by Esox are present in. the water.
C. The Gambusia react in this characteristic way only if they can see Esox or if chemical substances released by Esox are present in the water.
D. None of A, B, or C would be justified.

Exhibit 42.
Plant evolution, Science 1967

The following list describes five groups of plants.

LIST I

A. Plants with water-conducting tubes, leaves grow from underground stems, water necessary for reproduction, do not produce seeds.
B. Water-living plants, some single-celled, some many-celled, definite nuclei and chioroplasts.
C. Land plants, water-conducting tubes, produce seeds which lie unprotected on scale-like leaves.
D. Single-celled plants, no distinct nuclei or chioroplasts, live in water.
E. Small plants, definite nuclei and chioroplasts, no water conducting tubes, live in moist places.

The following features are characteristic of evolutionary development in plants:

(i) increase in size and complexity.
(ii) development of plant parts with specialized functions.
(iii) improved adaption to dry-land conditions.
(iv) increased efficiency of methods of survival.

Bearing these features in mind, rank the above groups of plants in ascending order of evolutionary development.

That is, next to Q.66 on the answer sheet, write the letter corresponding to the group that is at the lowest state of evolutionary development; next to Q.67 write the letter corresponding to the group that is at the second lowest stage of evolutionary development and so on to Q.70.

Exhibit 43. Fermentation, Science 1973 (see p.134)

The unit deals with the historical development of views about the fermentation process rather than current understanding. The observations of eight different scientists are recorded, and candidates have to see the relationship between the different views and recognise the way understanding of the phenomenon evolved over time.

Candidates had to draw inferences from pieces of information. They needed to see how much had been learned at a point in time, and what conclusions seem warranted at that time. They are given further evidence and asked to relate it to the different theories, and they are asked to make inferential attribution of ideas to different scientists.

In the final question they are asked to identify what the various observations show about the development of scientific theories. It is difficult to see the answer to this question. B and C can be eliminated because we have seen how accurate observation can lead to different conclusions. D is a reasonable statement in itself, but the passage is not concerned with the formation of theories, it is concerned with (as stated in A) the way ‘different theories may be put forward to explain a particular set of observations’. This question brings together the suggestions about how scientific knowledge develops as shown in the stimulus.

Exhibit 43.
Fermentation, Science 1973

During the eighteenth and nineteenth centuries various theories were put forward about the nature of the process of fermentation.

In 1837, **de la Tour** inspected drops of fermenting beer under a microscope and noticed yeast cells forming buds. The similarity of the process with sprouting seeds convinced him that the yeast cells were alive, and he decided that fermentation was the result of yeast growth and therefore part of a living process.

Pasteur's experiments on fermentation from 1857 to 1876 supported this theory. Investigating the cause of the fermenting brew going sour during its manufacture from beet sugar, he found that the alcoholic fermentation had been replaced in part by another kind of fermentation process which resulted in the conversion of sugar to lactic acid. Microscopic examination of the content of fermentation vats, in which lactic acid was being formed, disclosed that the round yeast cells, characteristic of alcoholic fermentation, had been replaced by rod-shaped cells much smaller than yeast cells. Successive transfers of minute amounts of these rod-shaped cells to fresh flasks of a sugar and water mixture always resulted in the production of lactic acid and an increase in the number of rod-shaped cells.

About 100 years prior to this work of Pasteur, **Lavoisier** had observed that, when sugar fermented, roughly equal masses of carbon dioxide and alcohol were formed, and be decided that fermentation was a simple chemical splitting of sugar into two other substances. This chemical theory of fermentation was supported by Wöhler, von Liebig, and Berzelius.

Wöhler's synthesis of urea (a constituent of urine from ordinary laboratory chemicals in 1828 provided strong evidence against the vital force theory. This theory held that only living things had the ability to build up unique organic (carbon-containing) molecules such as fats, sugars, proteins, urea, etc.

Von Liebig argued that fermentation was a process in which yeast cells died and decayed, forming an albuminous material which then rotted. The rotting material consisted of violently vibrating molecules; this vibration was passed on to the sugar which then fermented.

Berzelius, like von Liebig, regarded fermentation as a purely chemical process. In 1836 Berzelius published a paper in which he compared the action of yeast to that of a catalyst, i.e. he believed the actual yeast cells speeded up the formation of carbon dioxide and alcohol from sugar but did not take any biological part in the process.

About 1875, **Bernard** suggested that the yeast cells might contain an organic catalyst (enzyme), responsible for the fermentation of sugar, which might retain its activity even, after its extraction in soluble form from the yeast cells.

In 1897 **Buchner** ground up yeast and produced a juice which contained no intact yeast cells. He added sugar to this juice and found that the sugar fermented, forming alcohol and carbon dioxide. Buchner was able to show that some substance in the juice speeded up the fermentation process. This substance, which he called zymase, consists of enzymes, i.e. catalysts which occur in living organisms.

Questions 16-21 are based only on the information presented above.

16. A fermenting beet sugar brew which did not turn sour would probably contain

A. rod-shaped cells, but no round cells.
B. round cells, but no rod-shaped cells.*
C. both rod-shaped cells and round cells.
D. neither rod-shaped cells nor round cells.

17. 'The change of sugar to alcohol is a chemical process due to non-living substances, but these non-living substances are produced by living organisms.' This statement would most closely represent the view of

A. Lavoisier B. Berzelius C. de la Tour D. Buchner*

18. For a scientist working on fermentation during the period 1840 to 1850, the idea that yeast cells would be involved in fermentation is best described as

A. a fact which had been proved true by de la Tour's observations a few years earlier.
B. a hypothesis based on an observation, but one which had not been subjected to further experimental test.*
C. a theory which was incompatible with the theory of Berzelius.
D. a deduction which could have been made from the vital force theory.

19. Modern measurements show that the weight of one molecule of sugar is 180 atomic mass units (AMU), that of carbon dioxide 44 AMU, and alcohol 46 AMU; each sugar molecule can ferment to form two carbon dioxide and two alcohol molecules. These measurements provide evidence most closely related to the observations of

A. Lavoisier B. Berzelius C. Pasteur D. Buchner

20. 'The chemical act of fermentation is essentially a process correlative with a vital act, commencing and ceasing with the latter. There is never alcoholic fermentation, properly so called, without organization, development, and multiplication of cells or without the continued life of cells already formed.' Which of the following scientists is most likely to have written this passage?

A. Berzelius B. von Liebig C. Pasteur* D. Buchner

21. All of the information given in this unit illustrates that, in science,

A. different theories may be put forward to explain a particular set of observations.*
B. theories can be correctly deduced from observations.
C. if an observation can be found which supports a theory, the theory is thereby proved correct.
D. theories are generalized statements of observations.

Exhibit 44. Space flight conclusions, Science 1965 (see p.136)

In this unit candidate have to reason backwards to the basis or warrant for conclusions. This process of matching conclusions with observations is unusual. The issue is basic and concrete, and candidates have to inter-relate the information and make conceptual estimates. The unit involves a kind of logical decision making.

Exhibit 45. Radio-carbon, Science 1968 (see p.137)

This unit asks candidates to make inferences about assumptions underlying the conclusions of radio-carbon dating. Candidates have to understand how the process works, and reason about what is assumed in the first three questions. The last two questions involve calculations on the basis of the information. The unit as a whole is about the way radio-carbon dating is a matter of inferences based on assumptions.

Exhibit 46. Pecking order, Science 1969 (see p.139)

This curious and interesting piece of stimulus is both concrete and conceptual. It is science as systematic study and analysis of animal behaviour. One wonders if these things were known intuitively before they were systematically established in studies like the one reported.

This science is a matter clever observation and effective representation of data. The stimulus about a dominance hierarchy is conceptual and complicated, but the questions are rather disappointingly literal and procedural.

Exhibit 44.
Space flight conclusions, Science 1965

Read the passage below carefully and then the statements 26-29 following. These statements are all true conclusions based on the passage. For each conclusion circle on your answer sheet the letter of the sentence in the passage which must be used to reach it.

(K) In the space flight of a rocket a factor called the mass-ratio is most important.

(L) The take-off mass of a rocket is the sum of three quantities called the pay load, dead weight, and fuel load.

(M) The pay load is the mass of the compartments used for carrying explosives, scientific instruments and controls, or pilot and crew, and their contents.

(N) The dead weight is the total mass of the rocket structure plus all motors, fuel pumps and such like; and the fuel load is the mass of the fuel carried.

(0) After the fuel has been used up the pay load and dead weight form the remaining mass.

(P) The mass-ratio is the ratio of the take-off mass to the remaining mass.

(Q) Thus, if the take-off mass is 45 tons and the remaining mass is 15 tons, the mass-ratio is 3:1

(R) The higher the mass-ratio, the greater the fuel load and the greater the range of the rocket.

(S) When a single-stage rocket is fired it travels less than its own length during the first second.

(T) However, the velocity of a single-stage rocket increases rapidly, and it is about 40 ft./sec. at the end of the first second, 80 ft./sec. at the end of the second, 120 ft./sec. at the end of the third, and so on.

(U) Soon the velocities become very large, partly because the rocket is losing weight steadily as its fuel is consumed and partly because the effect of the motor increases as the rocket reaches the thinner layers of atmosphere where there is less resistance to its motion.

26. Half a minute after take-off a single-stage rocket has a velocity of the order of 2000 ft./sec.

27. If two rockets have the same dimensions and take-off mass, but one has a mass-ratio of 4:1 and the other a mass-ratio of 3:1, then the rocket with the 4:1 ratio will outdistance the other.

28. If improvements in scientific instruments result in lighter and smaller instruments it will be possible to use more instruments for a given pay load.

29. If in a rocket the fuel load accounts for more than half of the take-off mass, then the fuel load of this rocket exceeds the pay load and dead weight.

Exhibit 45.
Radio-carbon, Science 1968

Radio-carbon is a radio-active form of the element carbon, consisting of atoms slightly heavier (in the ratio 14 to 12) than ordinary carbon. It is usually called carbon-14 (or C^{14}) to distinguish it from ordinary carbon, carbon-12. A radio-active form of an element is unstable and as soon as it is formed begins to change into another element by a process we call radio-active decay. Carbon-14 atoms are being continuously formed in the earth's upper atmosphere from nitrogen atoms which are hit by cosmic rays from outer space, and as soon as they are formed they begin to change, in this case, back to nitrogen.

The half-life of C^{14} is about 5,700 years. This means that in 5,700 years half the original number of radio-carbon atoms change to nitrogen, and that in each succeeding period of 5,700 years, half those remaining change back to nitrogen. Moreover, it has been found that as fast as C^{14} atoms are transformed to nitrogen by radio-active decay, additional C^{14} is being created in the atmosphere by cosmic ray bombardment. Thus the concentration of C^{14} in the atmosphere remains constant.

As part of the growth process, plants absorb carbon dioxide containing C^{14} atoms along with ordinary carbon dioxide containing C^{12} atoms, and build their tissues out of it. When animals eat plants they absorb $C^{14}$and C^{12} in about the same proportion as they exist in the plants. As long as a plant or animal lives, the proportion of C^{14} to C^{12} in its tissues remains almost constant, as it is in the atmosphere, because the C^{14} lost by radio-active decay is replaced by C^{14} from the environment.

As soon as a plant or animal dies, however, the absorption of C^{14} ceases, and from then on the amount of this form of carbon in the remains of the plant or animal steadily decreases at a rate indicated by the half-life of 5,700 years.

So it is possible to determine when a plant or animal died by measuring the amount of C^{14} in its fossilized remains and comparing it with the amount normally present at death in similar plants or animals living today.

16. Which one of the following assumptions must be made in using this method of determining the age of a fossilized plant or animal?

A. No carbon has been lost from the fossilized plant or animal since its death occurred.
B. The amount of C^{14} in the fossilized plant or animal has remained constant since its death occurred.
C. The proportion of C^{14} to C^{12} has remained constant in the fossilized plant or animal since its death occurred.
D. The proportion of C^{14} to C^{12} in present day plants and animals is similar to what it was when the fossilized plant or animal died.

17. In this method of determining the age of a fossilized plant or animal, which one of the following assumptions must be made about the rate of bombardment of nitrogen atoms in the atmosphere by high energy cosmic rays?

A. The rate has been constant from the time when the animal or plant died to the present.
B. The rate is the same at present as it was when the plant or animal died.
C. The rate was much greater when the plant or animal died than it is at present.
D. The rate was much smaller when the plant or animal died than it is at present.

18. Which one of the following assumptions must be made in this method of determining the age of a fossilized plant or animal concerning the rate of radio-active decay of C^{14}, as indicated by the half-life?

A. The half-life is the same now as when the animal or plant died, but may have varied in the intervening period.
B. The half-life has remained constant from the time the animal died to the present.
C. The half-life is much greater now than when the plant or animal died.
D. The half-life is much smaller now than when the plant or animal died.

19. A piece of fossilized wood was found to have only half the proportion of C^{14} to C^{12} that a piece of living wood has. How old is the fossilized wood?

A. 2,850 years B. 5,700 years C. 11,400 years D. 17,100 years

20. A piece of fossilized bone was found to contain only one quarter the proportion of C^{14} to C^{12} as a piece of bone from a similar present day animal. How old is the fossilized bone?

A. 1,425 years B. 2,850 years C. 5,700 years D. 11,400 years E. 22,800 years

11

In pursuit of significant and sophisticated questions in the item writer's workshop

In Chapters 9 and 10, we have seen some exemplary multiple-choice stimulus and items, but the discussion of those examples is external and limited as it were. We might get closer to real excellence in item writing if we analyse and deconstruct in detail some material and the items written about it.

Good item writing begins with, and is always dependent on, good stimulus and effective analysis of that material. A good item writer can make good items out of very little, but even an expert item writer needs great stimulus to write great items.

There is some great stimulus for test writing on the next page. A candidate turns a page of the test and is confronted with a strange image. Is it an image? It seems to be a kind of matrix with two kinds of chickens, the confident white to the top right and the genuflecting dark to the bottom left. Whatever can it mean?

In a number of respects this *Pecking Order* unit (see Exhibit 46) can be a benchmark for judging the quality of test material. It is a great cross-curricular test stimulus because it:

- involves substantive thinking about an issue;
- is concrete, and conceptual and meaningful all at once;
- is genuinely challenging, provokes real questions and has potential for eliciting authentic scientific and quantitative thinking;
- is unfamiliar and unrelated to previous learning;
- is arresting and interesting; and
- teaches you something.

After the diagram there is an explicit prose explanation of how it works. A good deal is covered in the five items about the stimulus, and the items have the effect of leading the reader into understanding the diagram. You have to learn how to read it.

The first is a reflexive item about how you read the data. Items 18 to 20 involve the literal 'retrieval of information'. Item 21 is about the notion of the dominant hen, and Item 22 uses the concept of the dominant hen for an extrapolation exercise. This last item seems difficult at first, but when one grasps what it involves it is quite easy. This is characteristic of some great cross-curricular test items. Such questions at first seem hard, but if you understand the material, when you understand

the question they are quite easy. When you see that you only have to find which letter does not appear on the left of any pair (meaning it is not pecked by any hen) you have the answer to Item 22. M does not appear on the left in any of the six pairs, so it is the key.

Exhibit 46.
Pecking order, Science 1969
Items 17-23

	Y	B	V	R	G	YY	BB	VV	RR	GG	YB	BR
Y												
B	22											
V	8	29										
R	18	11	6									
G	11	21	11	12								
YY	30	7	6	21	8							
BB	10	12	3	8	15	30						
VV	12	17	27	6	3	19	8					
RR	17	26	12	11	10	17	3	13				
GG	6	16	7	26	8	6	12	26	6			
YB	11	7	2	17	12	13	11	18	8	21		
BR	21	6	16	3	15	8	12	20	12	6	27	

In any flock of domestic hens, each hen occupies a set place in an ascending hierarchy or 'ladder' of dominance, The hen on the lowest 'rung' of the ladder must submit to all the other hens. Higher up the ladder, a hen dominates those below it and submits to those above. The top hen dominates all the others. Dominance carries privileges. Thus the more dominant hens in a flock get first chance at the food trough, the dusting areas, the roost and the nest boxes, whilst the less dominant hens must give way and take what is left of the food and shelter.

The order of dominance in a flock is deduced from observations of the hens pecking one another. The chart opposite illustrates the peck-order of an experimental flock of twelve Rhode Island Red hens. Each hen in the flock is marked with either one or two colours; yellow (Y), blue (B), violet (V), red (R), and green (G). The hen with a yellow marking (column at far left) pecked all eleven of the other hens and was pecked by none of them. The number of times it pecked each one of them is indicated by the numbers in the columns. The hen with the blue marking (second column) pecked ten of the other hens but was pecked by the hen with a yellow marking. The hen with a blue and red marking pecked none of the other hens and was pecked by all. A hen submits to all hens which peck it, and dominates all hens which it may peck.

17. To find out how many hens pecked the hen labelled RR we

A. count the number of light-shaded hens in the 9th column.
B. add the numbers in the 9th column.
C. subtract the number of light-shaded hens in the 9th row from 12.
D. subtract the number of light-shaded hens in the 9th column from 12.*

18. How many hens could the hen labelled GG peck?

A. 2 B. 3 C. 9* D. 12

19. What was the label of the hen pecked by four other hens?

A. VV B. G* C. B D. RR

20. The hen labelled BB would not peck the hen labelled

A. BR B. RR C. YY* D. VV

21. The most dominant hen in the flock

A. pecks all the hens and is pecked by all of them.
B. pecks no hens and is pecked by none.
C. pecks all the hens and is pecked by none.*
D. peeks no hens and is pecked by all the others.

22. The letters H, J, K, L, M and N represent different birds in a flock of six hens.
The following instances of pecking were observed

K pecked H	H pecked J
L pecked N	J peeked N
M pecked L	L pecked K

23. The most dominant hen in the flock

A. was F. B. was L C. was M.* D. cannot be determined.

The candidate does have to understand how to read the diagram to answer these questions, but Items 17 to 22 about this stimulus must still be judged an honourable failure because the items did not do justice to the material. The items don't get much past information processing, and they are not sufficiently scientific or mathematical.

The stimulus invites many significant questions that are not asked in the unit, and the questions asked are not the really significant ones. Answering these questions means that the candidate can process the material but it doesn't mean that the candidate can see into the substance of the concept of pecking order. Seeing into the substance would involve the candidate identifying the key issues and seeing what is shown in the data about those key issues. Items 17 to 22 do not do this kind of exploration of the material, or allow the candidates to demonstrate that they can really deal with the issues. In that sense they are not the really significant questions about that material.

One suspects that the item writers knew that the *Pecking Order* unit did not 'work'. One suspects that the best items written about the material 'fell over' in trialling, in that they did not produce satisfactory trial statistics. It is a common frustration of the item writer's work that the pedestrian questions about some material 'work', but that the really good questions, those that are central and sophisticated do not work. It is a rough rule of thumb that the better the question (and that is not the same as the more difficult the question) the less likely it is to work because of the complexity of the thinking involved.

Multiple-choice questions about pecking order

In terms of this discussion, the distinction between the sophistication of an item and its difficulty is important. Items are not crude because they are easy and sophisticated because they are difficult. Closed and dichotomously scored items are sophisticated when they can get at things that are hard to target within the multiple-choice question constraints. (How can we ask a question about that issue in a multiple-choice?) In another sense test items are sophisticated when they require complicated awareness and reasoning from the candidates. Again sophistication is understood in this sense to be different from psychometric difficulty or working statistically. Straight vocabulary recall questions or shades of word meaning in context, for instance, can be very difficult and tend to reliably produce the required statistics, but they are not sophisticated items in the terms envisaged here.

The word 'sophisticated' as used here, is cognate with what the term 'higher-order' is usually used to mean in assessment parlance. There are problems with the term higher-order in that it has pseudo technical overtones (is there some natural order for kinds of question?), or it is assumed to imply questionable Piagetian assumptions about stages of development. The term higher-order also tends to coalesce with the notion of psychometric difficulty when they should be contradistinguished.

The testing of the understanding of irony, for instance, offers obvious examples of sophistication in test writing and in candidate performance. If a statement is ironic an item about it might offer the exact opposite of the statement as the key, and a paraphrase of the statement as a distracter. ('That's great!' which the reader understands to mean 'that is terrible!') Quite young candidates can sometimes recognise that the meaning of a statement is the opposite of what it seems to say literally. The same kind of thinking is the basis of an item, shown in Exhibit 47, about a cartoon in which one character says to another "It's all good" and which is glossed in the key as "I couldn't care less". Writing such items and getting them to work is difficult.

Exhibit 47.
Item from AST of 2010 based on a cartoon by William Haefel – *The New Yorker* cartoon of 2009

The cartoon suggests that the woman's comment is:

A. correct.
B. furious.
C. paranoid.
D. constructive.

"When you say 'It's all good' what you really mean is 'I don't care'."

It is often the case that the most sophisticated questions are much less likely to survive trialling because they are sophisticated. Alternatively, such questions (or what the item writer fondly thinks to be such questions) are shown by trialling to be eccentric. Such is the exploratory nature of sophisticated item writing.

What then are the central and sophisticated questions that might be asked about this pecking order material? What are the central questions that would have been asked about this material (they are clearly there to be seen in the stimulus) but that didn't survive the trialling?

The published questions remain at a kind of global level, and it is a striking omission that there are no surviving questions about the numbers on the subservient hens. What are those numbers? What do they show? What if anything do they tells us?

The numbers are clearly explained in the introductory text but there is nothing about them in the questions. The numbers in a cell, we are told, indicate the number of times a hen in a row is pecked by the hen at the top of that column.

Imagine asking a candidate the obvious, general question: What do we know about how the pecking order operates from the diagram? Or in more concrete terms, what do the numbers tell us about the pecking order? While these are obvious questions they are also the central, and hence, crucial questions.

It can sometimes seem that the best test items come at unusual angles and have counterintuitive answers. This is particularly true of the constrained multiple-choice question format. But it is usually the case that the obvious and general questions are those that really matter, and when the responses are to be scored by readers, it is the nature of the possible answers that is the issue. In such open constructed response items, it is the possible answers that are either simple or sophisticated.

Implicit in this big general question about the pecking order stimulus there are a number of more specific questions.

- While we know who pecks who and how many times, can we see any reason why one hen pecks a particular hen rather than another?
- What can we infer about the actual pecking of this or that hen?
- Is there a pattern to the instances of pecking?
- Do the hens go easy on those closer (metaphorically nearer) to them (are they allies?), and are they harder on those much lower down the pecking order than them? Or vice versa?

These questions are versions of the archetypal scientific questions about interpreting data.

- What is there to interpret in the data we have collected?
- How can we interpret this data?
- What conclusion or conclusions can we draw from this data?
- How should we go about analysing this data to draw conclusions?

These broad questions are quite simple in themselves, but, depending on the data, they can be easy or very difficult to answer. You have to do a significant amount of processing of the pecking order data to begin to answer them.

To frame the question precisely: what can we learn about the pecking order in this group from this data?

As they appear in the diagram, dealing with the numbers is difficult. It would help to extract the numbers into a table which can be put into a spreadsheet. The data is then easier to examine, and various calculations can be made.

One would want to add down the columns to see how often a hen pecks others, and one would want to add across the rows to see how often a hen is pecked. See the data and the results of these calculations in Table 17. One would then look for a pattern in the data by the usual first step of identifying the high values (hatched cells) and the low values (shaded cells).

Is there any pattern to be seen in these high and low values?

At first it looks like there might be something of a pattern in the highest and the lowest hens.

The dominant hen Y pecks the next highest B (22) and the lowest BR a lot (21). But Y also pecks the middle level YY most (30). Otherwise Y spreads the love around quite uniformly, although V (8) and GG (6) get off easily. It might seem that there is something of a bottom and top pattern to be explored further?

The lowest ranked BR (pecked most overall - 154) is pecked most by the hen immediately above (27), but also a significant amount by Y, the most dominant hen (21). BR is also pecked a significant amount by V (16) and G (15), but BR receives little unwanted attention from B (6), G (3) and GG (6).

Table 17. The number of times each hen pecks and is pecked

	Y	B	V	R	G	YY	BB	VV	RR	GG	YB	Pecked
B	22											22
V	8	29										37
R	18	11	6									35
G	11	21	11	12								55
YY	30	7	6	21	8							72
BB	10	12	3	8	15	30						78
VV	12	17	27	6	3	19	8					92
RR	17	26	12	11	10	17	3	13				109
GG	6	16	7	26	8	6	12	26	6			113
YB	11	7	2	17	12	13	18	18	8	21		127
BR	21	6	16	3	15	8	20	20	12	6	27	154
Pecking	166	152	90	104	71	93	61	77	26	27	27	

Overall, there is a near perfect pattern from top to bottom for being pecked (the sequence from 22 to 154 in the right end column), and a near perfect pattern from left to right for pecking (from 166 to 26) in the bottom row. But there is not a consistent pattern to the diagonals (varying between 6 and 29), and there is a mixture of high and low values in the middle range.

It can be concluded that in some respects the overall hierarchy is strict. There are no rebels, and there is no pecking of superior hens, but the operation of the hierarchy is not uniform after that. There is substantial variation in which hens lower in the hierarchy are pecked more or less by those above, and there doesn't seem to be much pattern to who a hen pecks below them and how often.

The data showing the number of individual pecks is complex and difficult to generalise. While the pecking order itself is strictly hierarchical overall, it is not uniform in operation.

In psychometric terms, the data is not a Guttman scale and it does not fit the Rasch model.

It would be interesting to see how far candidates could get in analysing the data in the chart. There are some simple and obvious things that can be said about the data, and there are some sophisticated things that can be said about it also.

The more difficult things could be made more accessible by allowing candidates to do some calculations in a spreadsheet, or they could be given the necessary calculations as in the chart and asked to interpret the results. The most sophisticated question is very simple in itself: 'What does this data show?'

In this sense, simple questions can be the most 'higher-order' in that candidates have to find what answer there is, if any, rather than assume there is an answer and search for it.

More sophisticated and difficult multiple-choice items could be written about the pecking order data than were presented in the test, as in the following.

> The counts of how often a hen pecks another hen show
>
> A. a clear pattern.
> B. little or no pattern.*
> C. a pattern at the top of the hierarchy but not at the bottom.
> D. a pattern at the bottom of the hierarchy but not at the top.

When trialled in 2015, 55.8 per cent of candidates chose the key of 'little or no pattern', and the item had a strong correlation of 0.38. Table 18 presents the results.

Table 18. Rasch analysis of Item 28

Item 28: item 28				Infit MNSQ = 0.96 Disc = 0.38	
Categories	A [0]	B [1]	C [0]	D [0]	missing
Count	22	86	29	17	1
Percent (%)	14.3	**55.8**	18.8	11.0	
Pt-Biserial	-0.28	**0.38**	-0.22	-0.02	
Mean Ability	-0.77	-0.06	-0.63	-0.34	0.22
StDev Ability	0.39	0.73	0.65	0.62	0.00

The fate of the crucial substantive question about the material was not as happy.

> An effect of the pecking order on the flock would be to
>
> A. cause conflict.
> B. structure conflict.
> C. minimise conflict.*
> D. make conflict predictable.

As Table 19 shows, while 25.2 per cent chose the intended answer of 'minimise conflict' (it is according to the original research the correct answer), the correlation of 0.16 is weak, and 38.7 per cent (with another weak positive correlation of 0.06) chose 'structure conflict'. Distracter B was too close to the key.

Table 19. Rasch analysis of Item 31

Item 31: item 31				Infit MNSQ = 1.11 Disc = 0.16	
Categories	A [0]	B [0]	C [1]	D [0]	missing
Count	17	60	39	39	0
Percent (%)	11.0	**38.7**	**25.2**	25.2	
Pt-Biserial	-0.18	**0.06**	**0.16**	-0.09	
Mean Ability	-0.68	-0.25	-0.10	-0.41	NA
StDev Ability	0.64	0.79	0.57	0.71	NA

Such global and open ended questions are difficult to make work as multiple-choice, but they are the most fundamental questions about such material.

- What is there to interpret in this data?
- How would you interpret this data?
- What conclusion can you draw from this data?

These questions are the kind of formulaic and archetypal open constructed response items referred to in Chapter 8.

Open-ended questions about pecking order

The pecking order stimulus was too interesting to use as a multiple-choice question so it was decided to use it as a short-answer question (or more accurately, a medium sized question). The pecking order stimulus was used as the sixth unit in the AST short response paper of 2015. There were two closed questions that were versions of the multiple-choice question, worth a mark each, and a broad and general question, worth four marks.

Question 10. Which hen is pecked by eight other birds?

In trialling, 57.1 per cent got this item correct as a multiple-choice question as shown in Table 20.

Table 20. Rasch analysis of Item 25

Item 25: item 25				Infit MNSQ = 0.99 Disc = 0.35	
Categories	1 [0]	2 [0]	3 [1]	4 [0]	missing
Count	12	31	88	23	1
Percent (%)	7.8	20.1	**57.1**	14.9	
Pt-Biserial	-0.13	-0.19	**0.35**	-0.17	
Mean Ability	-0.59	-0.58	-0.08	-0.60	-0.36
StDev Ability	0.35	0.64	0.73	0.67	0.00

As a short-answer question, 66.9 per cent of candidates answered Item 10 correctly, as shown in Table 21. These results are similar in difficulty when the trial population is usually about 12 per cent lower than the actual candidature. The short answer version is about the same difficulty as the multiple-choice.

Table 21. Rasch analysis of Item 10

Item 10: item 10			Infit MNSQ = 0.97 Disc = 0.42
Categories	0 [0]	1 [1]	missing
Count	954	1925	0
Percent (%)	33.1	**66.9**	
Pt-Biserial	-0.39	**0.39**	
Mean Ability	-0.13	0.29	NA

Question 11. Which hen is sixth in the pecking order?

As a multiple-choice question, the results shown in Table 22 suggest that 47.1 per cent answered correctly.

Table 22. Rasch analysis of Item 24

Item 24: item 24					Infit MNSQ = 1.04 Disc = 0.26
Categories	1 [0]	2 [1]	3 [0]	4 [0]	missing
Count	12	73	53	17	0
Percent (%)	7.7	47.1	34.2	11.0	
Pt-Biserial	-0.02	0.26	-0.32	0.08	
Mean Ability	-0.36	-0.09	-0.62	-0.14	NA
StDev Ability	0.59	0.74	0.68	0.42	NA

As a short-answer question, 55.3 per cent answered correctly in Table 23.

Table 23. Rasch analysis of Item 11

Item 11: item 11		Infit MNSQ = 1.01 Disc = 0.39	
Categories	0 [0]	2 [2]	missing
Count	1287	1592	0
Percent (%)	44.7	55.3	
Pt-Biserial	-0.35	0.35	
Mean Ability	-0.05	0.32	NA

Given that the trial population are about 12 per cent lower on average than the tested population, the performance for both the short-answer question versions was about the same as for the multiple-choice question versions.

Question 12. What does the chart show about the pecking order in this flock of hens?

There could be no equivalent to this open question in multiple-choice because there are a number of different answers, and the candidates had to decide what was important. The open item, Question 12 that followed the closed short-answer questions, along with the key features of the scoring are outlined in Table 24.

Table 24. Key features of Question 12 scoring

Score 3	There are three major conclusions to be drawn from the data. Candidates get 3 marks if they identify two of the major ideas. If they only identify 1 they get 1 or 2 depending on the accuracy of the description 1. Hens are pecked by those above (white) and peck those below (black) Hens do not peck those above them They peck all those below them to some extent 2. The higher in the hierarchy the less a hen is pecked overall. The further down the hierarchy the more a hen is pecked overall. 3. There is no pattern or order to how often a hen pecks individual hens below it. They don't for instance peck those closest more than those furthest, or vice versa.
Score 2	• They can be given 2 marks if they give a good version of one major idea. • To get 3 they must give more than one idea.
Score 1	• 1 idea poorly explained
Score 0	

The scoring identified three key points, and envisaged giving a mark for each point.

This open short-answer question was very challenging, with 43.0 per cent not scoring, and there was no attempt at an answer from 18.9 per cent of students. Only 4.9 per cent scored all three points, as Table 25 shows.

Table 25. Rasch analysis of Item 12

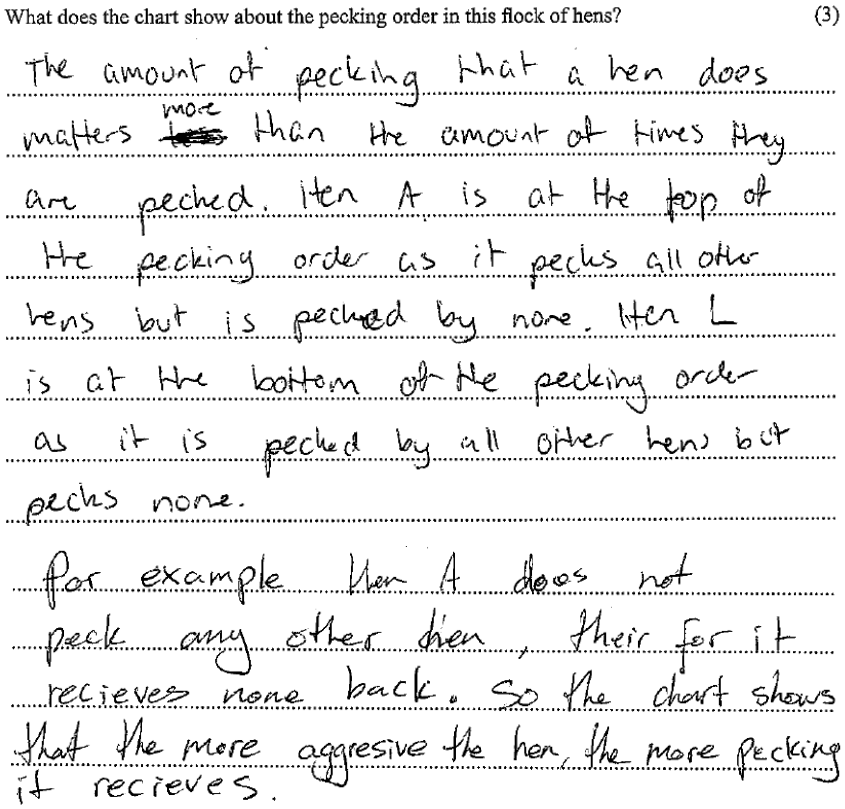
Item 12: item 12			Infit MNSQ = 0.97 Disc = 0.56	
Categories	0	1	2	3
Count	1239	956	543	141
Percent (%)	43.0	33.2	18.9	4.9
Pt-Biserial	-0.47	0.12	0.32	0.24
Mean Ability	-0.11	0.23	0.48	0.73

The third open short-answer question on the pecking order material was a higher-order question (and much more difficult) than the closed multiple-choice questions and short-answer questions in that candidates had to explore and analyse the data so as to identify the key features of it.

The first two closed short-answer questions brought to light a curious anomaly of closed multiple-choice and short-answer questions. There were 11 per cent of candidates who scored on both closed questions but could score nothing on the open third question.

The following response, shown in Figure 9, is by a candidate who got both closed short-answer question correct but the open question showed that the candidate did not understand the material.

Figure 9. Questions 12: Example response 1

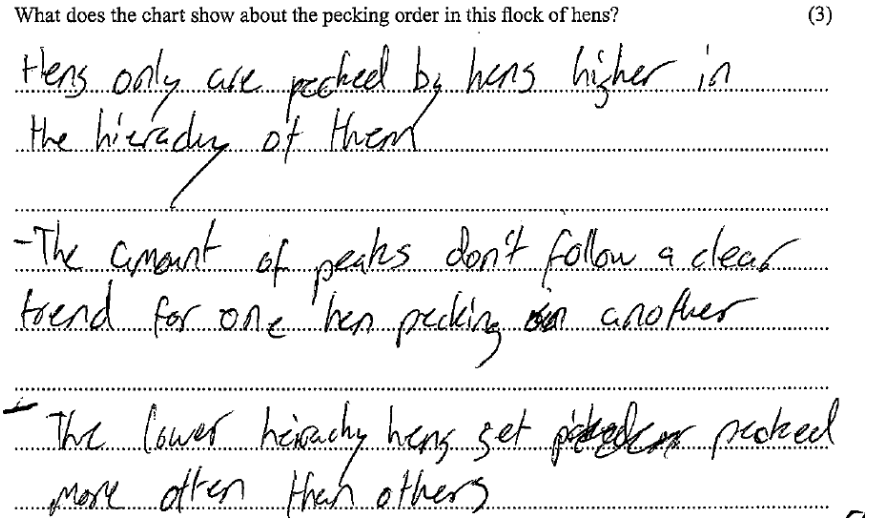

What does the chart show about the pecking order in this flock of hens? (3)

The amount of pecking that a hen does matters ~~less~~ more than the amount of times they are pecked. Hen A is at the top of the pecking order as it pecks all other hens but is pecked by none. Hen L is at the bottom of the pecking order as it is pecked by all other hens but pecks none.

For example Hen A does not peck any other hen, their for it recieves none back. So the chart shows that the more aggresive the hen, the more pecking it recieves.

On the other hand, some 36 words in Example 2, shown in Figure 10, gained full marks for the candidate.

Figure 10. Questions 12: Example response 2



The comparison of these multiple-choice questions, closed short-answer questions and an open short-answer question shows some interesting differences.

Literal, veridical and specific questions work much more readily in the multiple-choice format than broad, general and interpretive questions. There does not appear to be much difference between a multiple-choice question and a closed short-answer question version of the same question. On the other hand, an open short-answer question is much more demanding, and readily reveals the actual understanding of the candidate.

The questions about the pecking order material discussed here do not exhaust the possible approaches to that material. The more hypothetical or speculative question that failed as multiple-choice question about the result of the strict hierarchy on the flock would be a genuinely higher-order question. That question could be made simpler or more accessible by asking about the survival value of such a pattern of behaviour for a flock of hens. It would be interesting to see candidates would offer in response to such a questions.

12

In conclusion

We have seen that the Commonwealth Secondary Scholarship Examination (CSSE) was an unusually (and at that time, uniquely) broad and rich test of cognitive skills because it minimised the testing of knowledge and aimed to be a test of generic and cross-curricular thinking skills. The CSSE had four domain-related sub-tests, but it was a generic and cross-curricular conception of cognitive abilities in which the sub-tests were to be aggregated to give an overall score.

The breadth and richness of the CSSE is clear in comparison with the Scholastic Aptitude Test (SAT). It is hard to see a SAT of language knowledge, verbal puzzling, reading, and mathematics as adding up to a truly generic and cross-curricular construct.

The CSSE developed into a more specific and subject-oriented test in the Tertiary Entrance Examination Project (TEEP). The TEEP developed into a more generic and cross-curricular test in the Australian Scholastic Aptitude Test (ASAT). And as the research of Bell (1977) showed, as mentioned in Chapter 6, the ASAT was one test that could produce an overall score, and it could be partitioned into Quantitative and Verbal or (more accurately) Mathematics, science and technology (MST) and Humanities, arts and social sciences (HASS) sub-scores.

The notion of a cross-curricular test with MST and HASS sub-scores is a synthesis of the Spearmanian generalist and the Thurstonian specialist views of ability that is the red thread in the history of psychometric understanding of cognition. In psychometric terms, a cross-curricular test with MST and HASS sub-scores strikes a balance of coherence and differentiation.

A cross-curricular test with MST and HASS sub-scores is comprehensive and versatile. It can be used as a whole to compare the overall aggregate scores of students in one school with the aggregates of students in other schools, and it can be used to compare the performance of students in different subject areas with the performance of students in other subject areas.

It was the conceptualisation and production of such tests at ACER that turned item writing and test construction into an art.

References

Anderson, M. (1992). *Intelligence and Development: A Cognitive Theory*. Oxford, Blackwell.

Bell, R. C. (1977). A Psychometric Study of the ASAT Test (Series B). Western Australia, University of Western Australia, Research Unit in University Education.

Binet, A., & Simon, T. (1916). *The Development of Intelligence in Children: The Binet-Simon Scale* (No. 11). Williams & Wilkins Company.

Britton, J. (1970). *Language and Learning*. London, Allen Lane.

Britton, J. (1975). *The Development of Writing Abilities (11-18).* London, Schools Council.

Burt, C. (1940). *The Factors of the Mind: An Introduction to Factor Analysis in Psychology*. London, London University Press.

Carroll, J. B. (1982). The measurement of intelligence. In R. J. Sternberg (Ed.), *Handbook of Human Intelligence* (pp. 29-120). New York, Cambridge University Press.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York, Cambridge University Press.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book* (pp. xxi-xl). Boston, Houghton Mifflin.

Cattell, R. B. (1987). *Intelligence: Its Structure, Growth, and Action*. New York, North-Holland.

Ceci, S. J. (1990). *On Intelligence More or Less. A Bio-ecological Treatise on Intellectual Development*. Englewood Cliffs, N.J., Prentice Hall.

Clark, M. L., & Ennor, A.H. (1970). 21 July 1970, ACER archives, Series 42, Vol.327.

Connell, W. F. (1980). *The Australian Council for Educational Research 1930-80.* Melbourne, ACER.

Das, J. P., Kirby, J., & Jarman, R. F. (1975). Simultaneous and successive synthesis: An alternative model for cognitive abilities. *Psychological Bulletin*, 82(1), 87-103.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Massachusetts, MIT/Bradford Press.

Forrest, D. W. (1974). *Francis Galton: The Life and Times of a Victorian Genius*. New York, Taplinger.

Galton, F. (1883). *Inquiry into Human Faculty and its Development*. London, Macmillan.

Gardner, H. (1985). *Frames of Mind: The Theory of Multiple Intelligences*. New York, Basic Books.

Gardner, H., Kornhaber, M., & Wake, W. (1995). *Intelligence: Multiple Perspectives*. Harvard, Harvard University Press.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *Measurement of Writing Ability*. Research monograph, College Entrance Examination Board. 6.

Gosling, G. W. H. (1966). *Marking English Compositions*. Melbourne, ACER.

Gottfredson, L. (1998). The General Intelligence Factor. *Scientific American*, 9, 24-29.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, 24(1), 79-132.

Gould, S. J. (1996). *The Mismeasure of Man (Revised)*. London, Penguin Books.

Grant, A. & Withers, G. (1973). *From Testing to Teaching in the Humanities: Ideas and Resources for Use at the Upper Secondary School Level drawn from the Commonwealth Secondary Scholarship Examination 1971*. Hawthorn, Australian Council for Educational Research.

Haladyna, T. M. (1997). *Writing Test Items To Evaluate Higher Order Thinking*. Allyn & Bacon.

Hodgson, D. (1993). *The Mind Matters: Consciousness and Choice in a Quantum World*. Clarendon Press.

Horn, J. L. (1986). Intellectual ability concepts. *Advances in the Psychology of Human Intelligence*, 3, 35-77.

Horn, J. L. (1989). Models of intelligence. Intelligence: *Measurement, Theory, and Public Policy*, 29-73.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. *Encyclopedia of Human Intelligence, 1*, 443-451.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In P. Flanagan, ed. *Contemporary Intellectual Assessment. Theories, Tests and Issues*. New York, The Guilford Press.

Hunt, E. B. (1978). Mechanics of verbal ability. *Psychological Review,* 85, 109-130.

Hunt, E., Lunnenborg, C., & Lewis, T. (1975). What does it mean to be high verbal? *Cognitive Psychology*, 7(2) 194-227.

Jensen, A. R. (1998). The g factor and the design of education. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, Instruction, and Assessment: Theory into Practice*, (pp. 111–131). Lawrence Erlbaum.

Jensen, A.R., & Munro, E. (1979). Reaction time, movement time and intelligence. *Intelligence*, 3, 121–126.

Kaufman, A. S. (1990). *Assessing Adolescent and Adult Intelligence.* Boston, Allyn and Bacon.

Lawrence, I., Rigol, G. W., Van Essen, T., & Jackson, C. A. (2002). A Historical Perspective on the SAT® 1926-2001. Research Report No. 2002-7. *College Entrance Examination Board.*

Maling Keepes, J. B. R. (1973). *English and Its Assessment*. Melbourne, Australian Council for Educational Research.

Mayer, J. D., Salovey, & Caruso, D. (2000). Models of Emotional Intelligence. In R. Sternberg (Ed.) *Handbook of Intelligence*. Cambridge, Cambridge University Press.

McCurry, D. (2005). Notions of work-related skills and general abilities: The generic skills debate and the whole-school assessment of generic skills. Unplublished PhD Theses, Monash University Victoria.

Moffet, J. (1968). *Teaching the Universe of Discourse*. Boynton/Cook.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77-101.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence*, 3, 187-218.

Piaget, J. (1971). *The Psychology of Intelligence*. London, Routledge & Kegan.

Raven, J. C. (1936). Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. MSc Thesis, University of London.

Salovey, P., & Mayer, J. D. (1989). Emotional intelligence. *Imagination, Cognition & Personality, 9*(3), 185-211.

Spearman, C. E. (1923). *The Nature of 'Intelligence' and the Principles of Cognition*. London, Macmillan.

Spearman, C. E. (1927). *The Abilities of Man: Their Nature and Measurement*. London, Macmillan and Company.

Spearman, C. E. (1946). Theory of the general factor. *British Journal of Psychology*, 36, 117-131.

Spearritt, D. (1996). Carroll's model of cognitive abilities: educational implications. *International Journal of Educational Research*, 25(2), 107-198.

Sternberg, R. J. (1977). *Intelligence, Information Processing and Analogical Reasoning: The Componential Analysis of Human Abilities*. Hilesdale, NJ, Erlbaum.

Sternberg, R. J. (1989). *The Triarchic Mind: A New Theory of Human Intelligence*. New York, Penguin Books.

Sternberg, R. J. (1996). *Successful Intelligence: How Practical and Creative Intelligence Determine Success in Life*. New York, Simon & Schuster.

Thomson, J. D., & Keeves, J. P. (1974). *Some Studies in the Validity of the CSSE Tests*. Melbourne, Australian Council for Educational Research.

Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago, University of Chicago Press.

Vernon, P. E. (1973). *The Structure of Human Abilities*. London, Methuen.

Vogt, W. P. (1993). *Dictionary of Statistics and Methodology*. New York, Sage.

Vygotsky, L. S. (1986). *Thought and Language*. Cambridge Massachusetts, The MIT Press.

Wagner, R. K., & Sternberg, R. J. (1986). *Practical Intelligence: Nature and Origins of Competence in the Everyday World*. New York, Cambridge University Press.

Whitford, T. M. & Australian Council for Educational Research. (1967). *Educational objectives being tested in the Commonwealth Secondary Scholarship Examination*. Canberra. Commonwealth Govt. Printer.

Appendix 1

ACER General and Differential Ability Tests

ACER prepared the following general ability tests between 1930 and 1970s.

- The non-verbal test of general ability for Australia (1936-37).
- Intelligence tests (Grade 6) for NSW Department of Education (1937-48).
- Construction of intelligence tests for NSW Public Service Board (1940-54).
- General ability tests, prepared for Victorian Government scholarship examinations (1945-63).
- ACER intermediate D test (1946-51).
- ACER Junior A test (Grade 4), for NSW Department of Education (1946-47).
- ACER Junior B (verbal) general ability tests (1947- 51).
- ACER Jenkins non-verbal test: modification and standardization (1948-53).
- ACER Mechanical Reasoning Test (13-14 years) (1950-51).
- The general ability test for children aged 7 to 9 (1959-61).
- The general ability tests for the NSW Department of Labour and National Service (1962 – 66).

These tests were conventional intelligence test, and ACER has continued to produce such tests. The organisation currently sells the following.

- ACER Intermediate Tests: F And G
- ACER Advanced Test B40
- ACER Intermediate Tests: F And G
- ACER Advanced Test B40
- ACER Advanced Tests AL-AQ and BL-BQ
- ACER Applied Reading Test (ART)
- ACER Higher Tests PL-PQ and ML-MQ
- ACER Mechanical Reasoning Test
- ACER Test Of Employment Entry Mathematics (TEEM)
- ACER Test Of Reasoning Ability (TORA)
- ACER Tests Of Learning Ability (TOLA)
- ACER Word Knowledge Test: Forms E and F

Appendix 2

The CSSE Written Expression objectives of 1967

AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH EDUCATIONAL

OBJECTIVES BEING TESTED IN THE COMMONWEALTH SECONDARY SCHOLARSHIP EXAMINATION

Prepared in 1966 and amended in 1967 by

T. M. Whitford, Officer-in-charge Test Development,

in collaboration with the Chief Examiners in

Written Expression G. W. H. Gosling

Humanities K. J. Piper

Quantitative Thinking L. D. Blazely

Sciences N. L. Wilson (1964)
B. Rechter (1967)

and with the assistance of D. M. Bennett, R. T. Fitzgerald and J. P. Keeves

BY AUTHORITY A. J.ARTHUR
COMMONWEALTH GOVERNMENT PRINTER CANBERRA -1967

FOREWORD

Examination papers prepared by the Australian Council for Educational Research have been used in each year since 1964, to help to select those to whom the Commonwealth Government would offer a Secondary Scholarship. They were, when introduced, a type of examination new to Australia, and they have been both applauded and criticised.

For the sake therefore, both of those who prepare the examinations and those who use them, the A.C.E.R. thought it important to set down, in 1966, the background to their development and the rationale behind them. The result, with recent minor amendments, appears in the following pages.

While the C.S.S.E. papers are quite different from traditional general ability or intelligence tests, it will be clear that they are also unlike either the school examinations or the public examinations with which students are, likely to be familiar.

One obvious difference from the school or public examination is that the C.S.S.E. papers are not based on a specific course of study which all the examinees have followed, and which they have been deliberately taught. Instead, each paper attempts to test abilities which those at A.C.E.R. concerned with the examination believe will be developed by good teaching and interested learning, whatever the specific course of study. They are, we think, important abilities which are durable, and which should, if well developed, lead to success in later academic studies.

The C.S.S.E. does not directly test, as do school and public examinations, mastery of the knowledge and skills aimed at by particular courses of study. Some of the abilities which lead to this mastery are different from those tested in the Scholarship Examination, just as some of the abilities tested in the latter are different from those tested in the school examination. Consequently it is not to be expected that the two types of examination will rank students in exactly the same order of merit. An estimate of a candidate's potential success in later studies, if made by those familiar with the candidate's performance in school, should make a positive contribution to the data on which selection is based. The use of such an estimate in conjunction with the examination is therefore, in our view, a considerable improvement upon selection by examination alone.

The second obvious difference is that there are only four papers in the examination, and not one paper in each of the many subjects in the school curriculum at this level. When the A.C.E.R. was asked to suggest an appropriate competitive examination which would not limit a freedom that many schools cherish—the freedom to teach as they wish, what they wish—and which would not have bad effects on teaching and learning, the four papers referred to later were proposed. We believed they covered most of what was generally regarded as a good general education at the school level concerned and that the total testing time, while giving an adequate dispersion of marks, would not be too long for children or schools.

The test papers are different from normal examinations, therefore, and in our view for good reasons. Whether they are as effective in selection as other alternatives would be we do not know. By setting out the material which follows we expect to encourage informed consideration of the papers as modes of examining. Such consideration would involve the careful collection of data about both this examination and others which are taken by pupils in the course of their schooling. Upon the collection and wise interpretation of such data, improvements in the use and nature of this examination and others must depend.

Wm. C. RADFORD
Director Australian Council for Educational Research
May 1967

THE ORIGINS AND RATIONALE OF THE COMMONWEALTH SECONDARY SCHOLARSHIP EXAMINATION

Background to the C.S.S.E.

The C.S.S.E. papers were first proposed in 1964 to meet a specific need arising from the fact that, in the State of Victoria, there is no common external examination until the end of the Matriculation year. The Commonwealth Government required that the Commonwealth Secondary Scholarships should be awarded 'on relative merit as revealed in a test common to all competitors in that State' (Sir Robert Menzies: House of Representatives, 19 May 1964). To award scholarships on this competitive basis, the Victorian Government decided, after discussion between the then Director of Education and a number of educators, including the Director and Assistant Director of A.C.E.R., to use a proposal made by A.C.E.R. regarding a special examination, rather than reinstitute a subject-centred external examination. His proposal was that the examination for the award of the scholarship should consist of four two-hour papers, viz.:

- Written Expression,
- Quantitative Thinking,
- Comprehension and Interpretation in the Humanities,
- Comprehension and Interpretation in the Sciences,

and that the Victorian authorities should use the results from these, in any combination desired, to select scholarship winners. For the two-year scholarships, they decided to award scholarships in 1964 on the basis of the total (standardised) score on all four papers. In 1965 provision was made to incorporate school performance in the total mark. This was done by giving the school 'rank' of each candidate a mark equivalent to the score on a fifth paper. This use of a 'school performance' mark adds to the examination marks important elements not covered in the examination.

Though originally prepared for Victoria, the C.S.S.E. papers were also used by Western Australian and New South Wales authorities in 1964, and their use was extended to Tasmania in 1965, and to South Australia and Queensland in 1966. Technical scholarships were also awarded on the basis of these papers in South Australia, Western Australia and Queensland in 1966.

For 1967 the A.C.E.R. has developed papers which will be used in all states as in the previous year.

In February 1966 A.C.E.R. convened a meeting of representatives from each State to discuss and make recommendations on the nature of the papers and the objectives to be used as the basis of their construction.

The following statement represents in broad outline the rationale presented to and generally agreed upon by representatives attending the meeting.

Rationale

The rationale underlying the C.S.S.E. papers may be set but in the form of a number of propositions:

1. In the process of education (both formal and informal) a person develops skills and abilities as well as acquiring specific factual knowledge.
2. These developed abilities outlast much of the specific knowledge and are major objectives toward which a good teacher aims irrespective of subject course material taught in a particular school or State.
3. These developed abilities are important components of later academic success and it is therefore reasonable to use tests of such abilities to identify the most scholastically able students.

4. The most important of these developed abilities are considered to be:
    (a) Ability in the use of the mother-tongue, encompassing adequate vocabulary, recognition of correct and effective expression and the capacity to communicate correctly and fluently in writing.
    (b) Ability to read critically, to understand, to interpret, and to reason logically in the broad fields of the humanities and the sciences.
    (c) Ability to handle data presented in quantitative, symbolic or figural form.
5. It is possible to measure the development of these abilities through the medium of material which does not require the recall of specific information taught in subject courses.
6. Since tests based on the preceding propositions are concerned with the generalised outcomes of education and do not require teaching to a pre-determined 'subject-content' syllabus they do not restrict the freedom of the teacher. Further, such tests do not encourage memorisation of facts, formulae, procedures or mechanical application of standard formulae or procedures. Since they try to assess skills and abilities more important for later studies, any successful attempts at 'coaching' should result in relatively permanent educational gains as distinct from possibly temporary gains leading to success in the examination.

The general nature of the tests

The purpose of the awards being to encourage able students to stay at school longer and presumably to qualify themselves among other things for further education, the scholarship examination should aim to select those who, at the time of the award, display well developed scholastic abilities and appear most likely to be successful both at University Entrance or similar examinations and in later tertiary studies.

Though success in, say, a Matriculation examination may be a prerequisite for further education, results in such examinations do not always correlate particularly highly with success in later studies. A.C.E.R. considered therefore that success in a Matriculation or similar examination should not be regarded as a perfect criterion of subsequent scholastic achievement and should not, alone, determine the objectives of the tests to be constructed.

The view was taken that the objectives measured by the tests should, so far as possible, be of major importance and general applicability in further education. They should comprise abilities which the great majority of well-educated students wishing to pursue further studies could be expected to have developed by the time they took the examination.

It was decided to construct tests in Written Expression, Quantitative Thinking and Comprehension and Interpretation in the Sciences and the Humanities, as it was considered that these four broad areas were representative of major emphases in education. They were relevant both to further specialised studies and to the general education desirable for those not intending to undertake such studies after completing secondary education.

For example, Written Expression may be considered important not only in future language studies, but in all studies and other activities in which verbal communication plays a part. Quantitative Thinking abilities are relevant not only to further specialized studies in mathematics, but as a basis for appreciation and effective handling of certain data in geography, commerce, economics, medicine, etc.

The accent was placed throughout on the acquisition and use of general scholastic outcomes of good education rather than on highly specialised factual knowledge in specific subjects.

It is recognised that there are abilities not examined by the tests constructed in these four areas.

Firstly, students with marked abilities in foreign languages, music, art and craft and other subjects are not given specific opportunity to show these abilities. Provision of such opportunity would in A.C.E.R.'s view have required special 'content' tests, suitable only for those who had studied the subjects. If such special tests were provided for a portion only of the applicants,

(i) similar special tests would logically be needed for each subject studied at the level concerned, resulting in a virtual return to external subject examining, and the prescription of syllabuses, which certain State authorities wished to avoid.
(ii) inequities would have been introduced by the fact that no solution satisfactory to us in such an examination has yet been found to the technical problem of equating distributions of marks on different subjects which have been taken by different groups of students.

Secondly, even within the areas listed, there are developed abilities which cannot be adequately tested, for reasons of technical difficulty, at the present state of knowledge, or because of practical limitations imposed by the nature of large scale examining.

For example, the skills of listening and oral communication have not been included; with the exception of the Written Expression paper, no provision has been made for creative answering. Thirdly, though a comprehensive and acceptable list of measurable objectives may be drawn up for each broad area, the tests constructed will only *sample* these objectives, i.e. it is not practical to try to test everything in a two-hour paper.

OBJECTIVES UNDERLYING THE TESTS

General

A list of objectives to be measured by the tests has been drawn up for each of the four areas. These lists of objectives constitute A.C.E.R.'s present conception of major abilities which should have been developed by scholastically able students through formal and informal educational experiences over the years preceding the scholarship examination.

The objectives when translated into questions are of necessity concerned with the processing to some purpose of information presented in a particular form. That is, though the objectives are concerned primarily with the nature of the processing (thinking or production) necessary to achieve a particular purpose, consideration must also be given to the form in which the test material is presented.

As previously stated, these objectives relate, in the main, to abilities at a higher level than mere reproduction of specific knowledge or mechanical application of standard procedures or formulae.

In developing these lists and constructing the tests we have been strongly influenced by the principles and hierarchy of objectives set forth in Taxonomy of Educational Objectives Handbook 1: Cognitive Domain' edited by B. S. Bloom (1956, New York, David McKay and Co.).

In some instances we have drawn upon other statements, for example those prepared in connection with the I.E.A. Mathematics Project (International Evaluation of Educational Achievement, U.N.E.S.C.O. Institute for Education, Hamburg).

Though the tests are aimed at measuring developed abilities rather than specialised factual knowledge, a certain minimum knowledge and acquaintance with a given field must be assumed in order for communication of a problem to be possible. However, the extent of such assumed knowledge is limited. In the Quantitative Thinking paper for instance it is assumed only that candidates have a general familiarity with the mathematics covered in the first two years of secondary education.

Thus the 'non-specialist' is not precluded from attempting the questions in a particular paper because of lack of specialised knowledge. He may of course be less skilled than the 'specialist', who is often studying a subject because of special talent or interest, has greater familiarity with the area and may be able to call on techniques, concepts, etc., specifically taught in his field of specialisation. In other words, if education in a particular field is developing the abilities measured by these tests, then the intelligent

'specialist' should do better in a particular area than the equally intelligent ‘non-specialist'. On the other hand a very able 'non-specialist' may do as well as or better than many of the less scholastically able ‘specialists’ in a given area. The central point, however, is that the questions can be successfully attempted by candidates who have developed important and widely applicable capacities to understand, interpret, analyse, draw logical conclusions, etc., even though they may not be currently studying a particular subject.

The statements below relating to the four papers are regarded as working bases of test construction and not as definitive statements of the objectives of education in these areas. However, it is hoped that the statements prepared are in keeping with the general objectives of education which teachers aim to achieve.

WRITTEN EXPRESSION

Communicating via the written word is regarded as of major and crucial significance in most scholastic fields as well as having specific relevance to language studies.

The aim of the paper is to test attainment in this field by sampling the skills and abilities (including certain knowledge deemed prerequisite to their acquisition) which are developed by education in Written Expression during the primary and junior secondary years.

The area sampled may be conveniently subdivided as follows:

A. ***Specifics***, including

1. Grammar. Although a knowledge of any particular system of nomenclature or classification is not to be expected, a student should possess a knowledge of correct grammatical forms and structure. He should, for example, be able to give the form of a word which is correct in a given context, detect common errors and correct common errors.
2. Punctuation.
3. A knowledge of and the ability to recall appropriate words for use in a given context.
4. Word discrimination, e.g. given words similar or related in meaning to be able to assess relative suitability in a given context.
5. Discrimination between different sentences, paragraphs, etc., conveying the same idea, on the basis of correctness, felicity of expression, suitability of style to purpose, etc.
6. The arrangement of ideas, data, etc., into logical groupings and orders.

B. ***Integrated skills and abilities***

In the setting of composition topics, the Panel responsible for the Written Expression paper have been guided by the following:

1. Different types of composition (e.g. expository and narrative) test different skills. Any form of composition involves the integration of specific skills, not merely those necessary for all composition writing, but also skills specific to the particular type of writing task. Evidence from the 1964 and 1965 tests supports this conclusion.
2. Different types of composition are not regarded as being of equal importance. This, together with other factors, is borne in mind in deciding the frequency with which particular types of composition are set and the allocation of writing time to topics set.
3. It is considered legitimate to test in all areas where some degree of proficiency may be expected of students in the third last year of secondary education.
4. Each candidate may be required to display his or her ability in each type of writing task and options shall not be given.

5. Because options are not given, care must be taken to avoid topics where specialised or superior knowledge will confer an advantage upon particular students in terms of the content of their compositions.
6. As it is impossible to test all types of composition skill in any one year, as large a sample as possible should be examined within the available time.

Among the types of writing ability in which candidates for scholarships should have achieved a degree of proficiency are:

(i) A composition such as is written in a school subject other than English. Here, the data are known, and selection and arrangement constitute an important part of the task.
(ii) The ability to make outline notes on a given passage.
(iii) The writing of letters, e.g. to friends, relatives, a potential employer.
(iv) The ability to describe scenes, objects, etc.
(v) The ability to produce 'effects', e.g. to evoke an atmosphere.
(vi) The narration of experiences, either actual or imaginary.
(vii) The ability to present a reasoned argument for and/or against a particular point of view.
(viii) The ability to produce a composition based upon information presented in a form other than the written word, e.g. orally, pictorially or graphically.
(ix) Creative writing, e.g. telling -a -story with a degree of originality in theme or development.

Purposes and criteria

The measurement of abilities in the field of written expression involves consideration not only of classes of topic (i.e. (i) to (ix) above), but also appropriate criteria for assessment. These two aspects have been given detailed consideration by the examiners, and the following broad outline may prove of interest.

Purpose: A prime consideration in evaluating a written communication is to ascertain the writer's purpose, since this determines the criteria to be used in assessing the extent to which he has been successful. In the Written Expression paper, each topic set determines the purpose the writer (candidate) is required to fulfil through his composition. In some cases this purpose is rigidly defined; in others, some degree of freedom is given. Thus, of the classes of topic listed above, type (ii) is restrictive; type (v) relating to creation of an atmosphere in a certain specified context allows the writer some latitude in his approach to the task; while type (ix), being concerned with originality, gives the writer considerable freedom in theme or idea and approach.

Criteria: With respect to criteria it is essential to stress that, in each essay, the examiners are concerned with the *overall impression,* that is, with the demonstration of integrated skills appropriate to specific purposes. In arriving at an overall assessment of a particular composition, an examiner is required to give consideration to certain specific criteria appropriate to the task set. For certain topics, more weight is given to some criteria than to others, but there is no apportionment of marks to the separate characteristics of a candidate's composition.

The overall assessment may involve consideration of some or all of the following specific criteria:

- *Content:* Is the content comprising the essay, whether drawn from given data or from the candidate's own experience and knowledge, both *relevant* to the purpose of the composition and *adequate?*
- *Clarity:* Is the writer's meaning clear throughout?
- *Sequence:* Is the information presented in a sequence which is logical or appropriate to the development of the particular composition?
- *Unity:* Is the content of a paragraph homogeneous or does a candidate impose upon the material a unity appropriate to a paragraph? Is the paragraph characterised by appropriate 'linkage'? Where a composition consists of more than one paragraph, are the paragraphs so connected or presented that the composition forms an integrated whole?
- *Style:* Is the style suited to the writer's purpose? How well does he use language to create the effect at which he aims? Special attention may be given to the presence of any distinguishing features;

e.g. the use, where appropriate to the purpose, of anti-climax, irony or rhetorical questions, and the effectiveness of the opening and the conclusion.

The above list is not regarded as exhaustive, though it does contain criteria applicable to a wide range of writing tasks. Effective assessment of a particular task may require consideration of additional criteria and these will be incorporated into the essay marking guides as necessary.

QUANTITATIVE THINKING

Nature of the field covered

The objectives measured by this test are considered to be major abilities developed by mathematical education during the primary and junior secondary years. There are of course different views about the nature of mathematics and hence of mathematical education. Some people regard mathematics in terms of its practical uses and applications in a diversity of fields. Others emphasise its logical systems and the training in logical thought resulting from its study. A further group links mathematics with experience and interpretation of the real world and claims that mathematics assists in developing an understanding of the environment.

Despite variations in views about mathematics, there would probably be general agreement that the abilities developed by mathematical education in the schools are important for students undertaking further specialised study in that field and also for students planning to enter any of a number of other scholastic fields.

Further, it is possible to synthesise these different views in various ways to produce lists of objectives for school mathematics. Such lists or classifications of objectives form important systematic bases for organising instruction, and for developing tests of the products of mathematical education. The following classification of objectives has been adopted for the C.S.S.E. papers in Quantitative Thinking and has its roots in Bloom's Taxonomy and statements prepared in connection with the I.E.A. Mathematics Project.

Classification of objectives

The classification system can be summarised as follows:

(a) Ability to compute and manipulate numbers with skill and accuracy.
(b) Knowledge of definitions, notations, operations, etc.
(c) Ability to translate between verbal, symbolic, tabular, graphical, diagrammatic and pictorial material.
(d) Ability to read and comprehend new mathematical material.
(e) Ability to apply known concepts to new problems.
(f) Ability to analyse problems and to determine the relationships between their various parts.
(g) Ability to follow and construct proofs.
(h) Inventiveness - ability to reason creatively and to arrive at tentative intuitive solutions.

Notes on classification system

(a) The skills of arithmetical computations and arithmetic and algebraic manipulations are required to varying degrees throughout the study of mathematics. Likewise many of the solutions to the problems in the Quantitative Thinking examinations can only be reached after some computations have been performed. However, most of the figures are contrived so that, while knowledge of the computing process is important, the major difficulty of the question will not be due to the difficulty of the computation. There are one or two questions in which the main objective is to test the ability to compute accurately but in general it is considered that the important thing is to recognise the

order of magnitude of the answer. This category also includes the ability to make reasonable approximations and to apply reasonable computational checks to a solution in order to determine its general appropriateness.

(b) Because of the varied backgrounds of candidates, severe limits are imposed on the specific content knowledge that it is reasonable to assume as common to all candidates. No special knowledge of mathematics is required beyond the syllabus for the second year of secondary education in any State. In general, questions testing only knowledge have not been included, but of course some elementary mathematical knowledge is required in all questions. In some cases knowledge of special techniques or generalisations, etc., may lead to a more elegant solution but, in general, this possibility is avoided.

(c) Problems in mathematics can be presented in a variety of forms. The student is expected to be familiar with the techniques of the various basic methods of presenting mathematical relationships and ideas, to see equivalence between different ways of presenting the same information and to translate readily from one form to another. Included here are such abilities as giving a particular example to illustrate a generalisation; translating between three-dimensional situations and their two-dimensional representations; interpreting graphical representations of recorded data.

(d) This ability is concerned with the readiness with which students can comprehend new mathematical material and deal with it in terms of existing mathematical models or concepts. It includes the ability to see the implications of the new material and to extrapolate beyond the limits of what is given. This category has extra significance because it enables the influence of specialised knowledge to be minimised.

(e) The distinction between (e) and (f) may not be immediately obvious. In (e) it is intended that the relationships between the elements of the communication should be explicit and readily comprehensible. The problem lies in determining the operation or abstraction (theory, principle, etc.) that will enable the required result to be derived from the given information and also in carrying through the solution. For example, associating mathematical concepts with applications in community situations.

(f) In this category the relationships between the elements are not apparent and the problem is one of selecting, ordering and re-ordering so as to produce a consistent model. Included here are abilities such as being able to distinguish unstated assumptions in given information; formulating the necessary and/or sufficient conditions for a generalisation; distinguishing between relevant and irrelevant information for a particular purpose.

(g) This ability is concerned with the analytic and synthetic processes which are required in mathematical proof. While it is relatively simple to test the analytic aspect of proof, objective tests such as the Commonwealth Secondary Scholarship Examination, do not lend themselves to the testing of the ability to synthesise a proof. Attempts are being made to devise a satisfactory method of testing this latter ability.

(h) It is evident to the examiners that many students arrive at answers by intuitive methods in many questions where analytic solutions are possible. Specific testing of intuitive or creative problem solving, however, presents considerable difficulty and although A.C.E.R. has developed experimental material related to this ability, no satisfactory testing method compatible with the general nature of large scale examining has been devised.

It should be made clear at this point that a classification of this type does not imply that any task within the area can be neatly placed in one of the above classes. It is apparent than an order of complexity could be established between some of the classes listed. If a task involves one of the more complex abilities within such a hierarchy then it is likely that other less complex abilities are also involved.

The abilities listed cannot be measured in a neutral setting. The materials presented to a student have to be in a certain form and presented in a particular context. In the absence of any real evidence it has to be assumed that ability, context and form are not independent and the latter two aspects have to be considered in preparing a systematic basis for test development. The form in which the material is presented can be easily categorised into the five classes: words, symbols, graphs, tables and diagrams. However, the classification of topic area or context presents certain problems as a number of possible

systems of classification are available. The traditional division into arithmetic, algebra and geometry has been rejected as not being particularly useful and also as not in keeping with trends towards unification in mathematics curriculum development. Consequently it has been decided to use the following breakdown related to the source of the material - social sciences, economics, physical and biological sciences and pure mathematics. There are some abilities, particularly those related to communication skills that cannot be satisfactorily tested by using the present examination format which includes only objective questions. However, it is considered that the present style of papers samples the listed objectives extensively, and that the paper is an effective and efficient way of testing most of the major objectives in this field.

COMPREHENSION AND INTERPRETATION IN THE SCIENCES

A view of the nature of science

Although a learned exposition of the nature of a science would be inappropriate here, some consideration of the question seems necessary as an orientation to the statement of objectives. The following comments, while brief and oversimplified, present a view of the nature of science which has been found useful in analysing objectives of science education and in constructing the C.S.S.E. paper.

Science is concerned with the relationships between aspects of the environment. Knowledge of such relationships enables us to describe accurately the outcome of interactions within the environment, and thus to some extent to predict, modify or control the outcomes of such interactions.

A science in its elementary stages may be concerned largely with classification and identification of a class of phenomena in terms of some aspect which:

(a) describes the class;
(b) differentiates the class from other classes;
(c) seems to be relevant to the interactions associated with the phenomena.

In the physical sciences, and to a lesser extent in the other sciences, the properties which determine such a classification are quantified, definite experiments are carried out to determine the relationships between these quantified properties, and some model is developed with similar property relationships.

Two aspects of the formal development of a science may be considered:

(a) Empirical generalisation, in which a law, in universal or statistical form, expresses connections between directly observable phenomena.
(b) Theory formation - the empirical generalisations of (a) may be derived from more comprehensive laws (expressed in terms of hypothetical constructs).

Generalisations in science involve the use of models. For our purposes, a model is regarded essentially as a statement of relationships between constructs. The construct may refer to some directly observable or measurable property of the physical or biological environment, or may itself be defined by a set of properties.

The models (laws and hypotheses and theories) of science are more or less consistent with observable phenomena. That is, the relationships of the model conform more or less to the relationships between certain properties of the environment.

The production of, prediction from and verification of suitable scientific models of the environment presuppose certain problem solving skills (amongst other skills and abilities) in the worker in the field. These problem solving skills should be present, although probably in a less developed form, in the student of science, and in still less developed form in an educated 'layman' interested in science.

The test developed by A.C.E.R. attempts to measure the extent to which these problem solving skills have been developed by students presenting for the examination.

Problem solving skills

The problem solving skills referred to above may be conveniently classified as follows:

Classification skills

1. Identifying or discovering the property, or group of properties, which classify the observable object or phenomenon (e.g. recognising that it is density which will determine whether a body will float).
2. Classifying an observable object or phenomenon in terms of the properties which predict its behaviour (e.g., given the relevant information about density, decide whether a body will float).

Translation skills

3. Recognising correspondences between constructs and observables (e.g. realising the correspondence between pressure and a dial reading, or balloon size).
4. Recognising different forms of the construct (e.g. realising that pressure is the same as force per unit area).
5. Recognising that relationships within the model correspond to relationships within the observable environment.
6. Recognising different forms of the relationship (e.g. recognising a direct proportionality relationship whether presented in symbolic or verbal form).

Extrapolation skills

7. Predicting necessary or probable consequences of the model.
8. Testing the consistency of a model with observable phenomena.

Application skills

9. Recognising correspondences between the model and new phenomena, where the correspondence is not explicitly referred to (e.g. recognising that a general law will apply to a particular situation).
10. Recognising correspondences between different statements of the model where the correspondence is not explicitly referred to (e.g. recognising the correspondences between a diagrammatic and a symbolic statement of a chemical system or biological process).

Synthesis skills

11. Producing a suitable model to produce a suitable relationship between suitable constructs.

Subject areas and form of questions

Though the accent is on measuring the extent of attainment of the objectives or developed abilities outlined above, the preparation and selection of questions has also involved consideration of subject areas and forms of communication.

Subject Areas: Any study satisfying the requirements of the 'nature of a science' as described above may be considered as a suitable subject area. Questions have been prepared mainly in the following contexts: Physical Sciences (Physics, Chemistry and their applications); Earth Sciences and Astronomy; Biological Sciences.

Form of communication

The relationships - implicit or explicit - between aspects of the environment are presented to the student in terms of a printed communication which may be verbal, symbolic, diagrammatic or graphical in form.

Similarly, the constructs (properties, phenomena, classes) between which the relationship mediates have been expressed in various forms e.g. verbal, symbolic, quantitative or diagrammatic.

Much advanced scientific thinking utilises symbolic relationships. While the mathematical tools available to students in the third last year of secondary education are limited, it is assumed that candidates comprehend the proportionality relationships represented by simple algebraic formulae.

The complexity of the problem solving situations

The problem solving skills listed above are probably hierarchical in the sense that the application and synthesis problems, which involve implicitly rather than explicitly stated relationships, require more highly developed cognitive ability for their solution. It may also be that relationships expressed in symbolic or graphical form are more difficult for many students than are similar relationships if expressed verbally.

There are, in addition, two other factors affecting complexity which should be mentioned. These relate to:

(a) The number of variables and relationships involved in the statement of the model. For example a statement of the form F=k a, where both 'k' and 'a' are variables, is less complex if only 'a' is a variable.
(b) The 'noise', or material irrelevant to the solution of the problem which may be included in the stimulus material, and to a lesser extent in the alternatives of a multiple-choice question. This 'noise' may affect the student's ability to 'retrieve' the relevant information at the required time in order to solve the problem, or his ability to match the relevant relationship with the correct alternative.

Restrictions on questions

There are doubtless abilities required by workers in the fields of science which have not been included in the nicely contained system outlined above. It is stressed that the system is not necessarily all-inclusive, nor will the examination necessarily be restricted to questions which fit this particular model of science-reasoning abilities. In general, however, questions will conform to the overall pattern that has been outlined.

COMPREHENSION AND INTERPRETATION IN THE HUMANITIES

The term Humanities as used in relation to the Commonwealth Secondary Scholarship Examination covers a variety of subject areas in the social sciences and the traditional humanities, e.g. Literature, History, Geography and Social Studies, though the accent throughout is on the thinking processes involved rather than on particular subject matter.

Though it is recognised that languages and the arts should also be included in the subject areas, satisfactory testing in these two fields is not considered practicable in the present type of large-scale examining programme (see The general nature of the tests, pages 2 to 4). However this issue is not regarded as closed and A.C.E.R. is at present investigating the extent to which skills in these two particular areas are or may be tested in this type of examination.

The aim of the paper is to test a number of the major abilities involved in or developed by studies in the humanities. Although the tests are not concerned with specific subject knowledge, many of the questions call upon a background of general understanding and appreciation of human factors in the areas studied as well as on cognitive abilities, such as logical reasoning, which have more general application in fields other than the humanities.

The paper does not test all the abilities associated with the humanities, nor does it specifically attempt to assess attitudes and other affective outcomes of education in these fields. The abilities listed are not

regarded as being in a necessarily hierarchical order of complexity, importance, or sequence of application to a problem, and in particular problems some of them will be of more relevance than others. It is also essential to remember that some overlap and interaction occur between the separate elements and that any particular problem will cut across such artificial classifications and require for its solution many, if not all, of the abilities listed. Since the accent is on generalised abilities rather than on specific subject matter, the test material used in any one paper will sample but not necessarily cover the range of subject areas listed.

Question situations in the field of Humanities may be presented in various forms, e.g. the written word (prose, verse), pictures, diagrams, maps, graphs, etc., but in much of the material so far produced the written word is the major form used.

In a number of fields subsumed for our purposes under the general heading 'Humanities', problems may be reduced to the question 'What relationships exist within, or what conclusions, may reasonably be drawn from, the available information?'

The relationships detected or conclusions drawn by a worker in this field depend upon his capacity to understand, interpret, assess and synthesise the information given. This capacity as analysed below forms the basis of the material set in the C.S.S.E. Humanities paper.

List of abilities

1. *Comprehension*

The requirement here is understanding of both the explicit and implicit meaning of information given, i.e. the denotations and connotations of words separately and when grouped into statements. This covers:

(i) understanding the information as *explicitly* stated;
(ii) understanding *implications;*
  (a) recognition of *connotations* of words and statements;
  (b) deducing *necessary* logical consequences of the information given;
  (c) deducing *likely* consequences or associations of the information given, e.g. being able to extrapolate to the extent of predicting likely consequences not explicitly given.

2. *Critical assessment of material and source*

The requirement here is critical examination of all available information in order to make some assessment of statements or persons. The statements (i.e. the material presented) may need to be assessed to judge their significance or usefulness for some purpose.

Assessment of a person (e.g. the author of a statement, or some person about whom information is presented) may be necessary in order to make predictions about his likely views, actions, etc.

Such assessments involve the checking of relationships in the material presented and the application to the material of criteria previously learned, stated, or developed from the information.

(a) *Assessment of the material* is a complex process which may require critical examination of the material (i) as it stands, and in relation to (ii) its source and (iii) other information presented or previously known. These three are closely inter-related in assessment of material.
  (i) *The material as it stands* may be considered in terms of two related aspects:
    - *Logical content of the material:* Application of criteria relating to argument, e.g. distinguishing between fact and opinion or verifiable and non-verifiable data; identifying explicit or implicit assumptions; judging appropriateness and adequacy of evidence for a purpose; recognising contradictions or inconsistencies within statements; detecting logical fallacies in argument; detecting unwarranted extension of argument beyond the limits of assumptions or evidence presented, etc.

- *Form or style of the material:* Recognition that form or style of presentation may provide a part-basis for evaluation of material for a given purpose. This may involve detecting bias, persuasive or propaganda techniques, suppression or selection of evidence, etc.

(ii) *The source:* Information or deduction about source or authorship may be relevant to assessing the significance of the material for a given purpose. This assessment may involve criteria associated with the known or probable characteristics of the author; deciding whether the source may be regarded as authoritative (knowledgeable and responsible); considering who wrote the communication, when and under what conditions; being alert to the likelihood that author's bias, motives, purposes, etc., would influence certain statements he might make.

(iii) *Other evidence* presented may be important in evaluating statements for a given purpose. This evaluation may involve determining the nature and extent of relationships between the statement being assessed and any other given fact or statement (which latter may itself need to be assessed in terms of (i) and (ii) above); arising from application of these criteria, determining what additional information is required to substantiate or disprove an argument, etc.

(b) *Assessment of persons* (authors, sources, etc.) is relevant to the prediction from available information, of likely actions, beliefs, etc., of a person. Such predictions may arise from critical assessment of:

(i) statements about the person (possibly considered in relation to (ii) and (iii);

(ii) the person's own statements about a topic, considered alone or in relation to

(iii) other evidence about the same or a related topic. Such assessments may involve analyses and comparisons similar to those required in (a) above.

3. *Drawing conclusions* about the topic dealt with in the material presented. This may or may not require prior critical assessment of the material. It may, for instance, be permissible to accept the material presented and to proceed to draw conclusions about the topic which would necessarily or probably follow if the material were unchallenged. Included here are: making generalisations; detecting various kinds of relationship e.g. causal, chronological, concurrent, etc.; perception of trends; classification of material into given categories; selecting and re-arranging data to arrive at a conclusion which is new or different from one stated or implied in the material; formulation of hypotheses to account for facts.

4. *Appreciation:* Abilities under this heading are primarily concerned with aesthetic appreciation and encompass both sensitivity to and critical assessment of literature. Among these abilities are the following: Awareness of style and stylistic differences; awareness of effectiveness of different statements of an idea; sensitivity of response to literary and artistic qualities in the material presented; evaluation of various forms of expression; judgment of quality of expression; critical judgment of form as distinct from content; awareness and evaluation of the whole as well as the component parts.

5. *Creativity:* Creative elements are no doubt involved, for example, in the abilities described in 3 above, where a problem may be solved by an insightful synthesis of separate ideas or pieces of information into a 'new' concept of a situation. However, direct testing of creative or imaginative production has not yet been found practicable in the Humanities paper of the Commonwealth Secondary Scholarship Examination.

Appendix 3

A revised view of the CSSE Written Expression paper and its objectives 1972

EDUCATIONAL OBJECTIVES being tested in the CSSE
A revised view of the Written Expression paper and its objectives

1972

The statement of objectives for the CSSE Written Expression paper, made in 1967, is probably, of the four statements, the one most in need of revision in terms of changes in attitude since then. The other three papers are still totally multiple- choice, whereas the written expression paper no longer tests the specifics of grammar and word usage and is totally given over to eliciting pieces of original writing. In the terms of the 1967 statement, 'communication via the written word' is still regarded as of 'major significance in most scholastic fields', but the means by which language competence is tested have undergone significant alterations.

While the 'specifics' and 'integrated skills and abilities' mentioned in the 1967 statement still, to some extent, apply to the examining in 1972, we would not now wish to sample such areas directly. Perhaps the easiest way of showing both their continuing relevance and their present status in the assessment of writing ability is to indicate the way in which essays are presently marked. We use a team of 200 markers, and each is asked to use the criteria which they would normally use, as teachers, in deciding one piece of writing was better than another. Given that each paper is randomly assigned to eight members of the team, the variety of criteria employed in judging it is likely to be wide. But since all markers are trained, qualified and experienced teachers of English, some account of the specific and integrated skills will be taken by each marker. Some markers are likely to include specifics such as grammar, punctuation and word usage high in their sets of criteria; others are likely to make much less of such specifics and concentrate on integrated skills, such as narrative power, ability to write a reasoned argument or expression which has a high degree of originality in form or subject matter. Assigning eight markers to each candidate's work ensures that a cross-section of opinion among English teachers as to what constitutes 'good' writing is brought to bear on each paper.

Since 1967, the weight of responsibility for the setting of standards of 'goods writing' has shifted from the Chief Examiner to individual markers. Each person is concerned, in making judgments, with the 'overall impression' as before; that is, with the demonstration of those integrated skills mentioned above. No longer, however, does the Chief Examiner define, or set limits to, the acceptability or appropriateness of particular skills to specific purposes. No marker is now required to give consideration to any specific criteria 'appropriate' to the task set, unless they are his own criteria. In the circumstances of double marking 400 000 essays in three weeks, no criteria can reasonably be specified because their application cannot be checked, and such checking would be against the professional spirit we attempt to encourage among

markers. We rely on the training, experience and professional integrity of each marker to make his or her mark valid. Investigations of the reliability of the marking are made from time to time; they give us reason to believe that candidate's writing is treated sympathetically and fairly and that total scores are a reliable and valid reflection of the criteria currently used by Australian teachers of English in the assessment of writing ability.

In setting the topics, the panel responsible are now guided by somewhat different assumptions from those published in 1967. Evidence more recent than that quoted (p.6) now suggests that two examples of a particular type of composition (e.g. two narrative topics) are likely to produce writing and results which differ as widely as two tasks which call on different types of writing (e.g. a narrative topic as opposed to a descriptive one). Further to this, one of the chief criteria in choosing topics is to select ones which have worked well in trial testing at the appropriate grade level. Options are still not given, and care is still taken to avoid topics where specialized or superior knowledge will confer an advantage on particular pupils. These four aspects lead us to select topics which give all students a chance to write, and to choose their own mode of discourse for the topic. A particular topic will produce essays in many modes: a single task will bring forth arguments, conversations, descriptions and stories in prose or poetry, and sometimes dramatic form. Therefore in selection of the four topics for any particular year care is taken to provide as wide a range as possible of stimulus material and themes - the ways in which a theme might be taken up are left for the student to review and choose from. Markers are asked not to exercise too strict a definition of relevance when they come to mark the essays, in order to give even the weakest student a chance to score well for his or her writing. Evidence suggests that pupils operate well in this situation: less than two per cent of candidates do not complete the paper (i.e. write four essays, each of which is assessable), and we can be sure that some of this number fail to complete it because they run out of time, not because they lack motivation to write about a particular topic.

We would not therefore now wish to suggest that a candidate should have achieved a degree of proficiency in any particular types of writing ability (such as those specified under sections 6i to 6ix on pages 6 and 7). As English teachers, we might hope that all candidates would have had contact with these (and other) forms of writing, but any sense that experience and proficiency are prerequisites for candidature should now be removed. Sometimes the wording of a particular topic will suggest or imply an appropriate form or mode for response, but if a pupil ignores or fails to see such implications he or she is not penalised. The main purpose of each topic is to get pupils to write, as well as they possibly can, especially since Written Expression is the only paper of the four which permits the candidates to respond in original and extended forms. We have reason to believe that the removal of all rigid definitions of how they shall write has encouraged this. At first sight, such a testing technique might seem to place the very imaginative child at an advantage, but a survey of the essays which receive high marks reveals that a large proportion of these are simple, straightforward expressions of opinion or statements of fact, stories or descriptive prose, alongside the highly imaginative, creative pieces of work.

Many of the comments made on pp.7 and 8 under 'Purposes and criteria' still stand, as general indications of how marking is carried out. Criteria such as content, clarity, sequence, unity and style will, as stated above, inform to varying degrees the judgments made by markers. However, we wish to point out that essay marking guides, prepared by a chief examiner or panel, are no longer used. Reliable assessment can be achieved in essay examinations without this sort of external determination of purposes and criteria if the marking team is carefully chosen and the distribution of scores is controlled statistically. The markers are asked to award grades in a distribution which very roughly accords with the normal curve: checks encourage us to believe that this technique, in conjunction with all the other setting and marking aspects of the scheme outlined above, allows a free expression of ideas and opinions in candidates' writing, encourages the demonstration of integrated writing skills, and ensures that the total scores achieved for four pieces of writing are a reliable consensus of opinion about the worth of the writing thus summoned from the candidates.

Appendix 4

The ACT Scaling Test and School-based Assessment

In the Australian Capital Territory senior secondary education follows a system of externally-moderated school-based assessment. In this system, students are assessed by their teachers who are the people most familiar with students' achievement in a course. The basis for judging a student's achievements in the course are the criteria and standards specified in the relevant Course Framework, but the teachers' judgments must be able to stand up to scrutiny by expert practitioners external to the particular school. This process of quality assurance is referred to as moderation, and involves both the Board and the school.

The Purpose of Moderation

Moderation is a broad system designed and implemented by the ACT Board of Senior Secondary Studies (ACT BSSS). It ensures system-wide comparability of grades and scores on the ACT Year 12 Certificate. It is designed to:

- provide comparability in the system of school-based assessment;
- form the basis for valid and reliable assessment in senior secondary schools;
- involve the Board and schools in cooperation and partnership; and
- maintain the quality of school-based assessment and the credibility, validity and acceptability of Board certificates.

To ensure that the information appearing on students' Year 12 Certificates is comparable and consistent both within and between schools in the ACT Senior Secondary System, the Board undertakes a system-wide external moderation process.

The ACT Scaling Test is designed by the Australian Council for Educational Research. It aims to measure a range of general skills considered to be relevant to success in university studies. All students seeking admission to university must sit this test in order to gain a University Admission Index (UAI). Student scores in the test provide the basis for scaling course scores across colleges using the Other Course Score Scaling (OCS) method. The test is not used to directly adjust a student's own course scores. Fuller details are included in the booklet *What's the AST?* available in schools and from the Board Secretariat.

For higher education consideration, the student must sit the ACT Scaling Test (AST), which is used to adjust scores awarded in T courses to achieve comparability of students' achievement across colleges and courses.

http://www.bsss.act.edu.au/
http://www.bsss.act.edu.au/year_11_and_12/act_scaling_test

Appendix 5

ACT Scaling Test Specification 2011

1. ACT SCALING TEST REQUIREMENTS

1.1 BACKGROUND

1.1.1. The ACT Board of Senior Secondary Studies (BSSS) is a statutory authority responsible for the certification of senior secondary school studies in government and non-government schools in the Australian Capital Territory.

1.1.2. The principal functions of the BSSS are to:

(a) accredit or register courses taught by recognised educational institutions;
(b) approve, consistently with national agreements, recognised educational institutions for teaching vocational education courses;
(c) establish guidelines for the development of courses by the Board or by a recognised educational institution;
(d) establish principles and procedures for the assessment of attainments of students and the moderation of these assessments; and
(e) provide to people who have undertaken courses, or units of courses, certificates and transcripts of their attainments, including the ACT Year 12 Certificate.

1. MULTIPLE CHOICE SPECIFICATIONS

1.1 GENERAL DESCRIPTION

1.1.1. The Multiple Choice component of the AST is to consist of a two and a quarter hour paper (135 minutes) of 80 questions. It is intended for administration to Year 12 ACT and affiliated overseas schools'. Each question will have a structure of a stem and four (4) options.

1.1.2. Questions are to be grouped in units, based on stimulus material in a variety of forms. Any specific information required to answer the questions should be contained in this stimulus material, pitched at a level of vocabulary knowledge assumed to be common to students in Year 12.

1.1.3. The questions should be designed to measure, in the main, the abilities of comprehension, interpretation and reasoning, across as many curriculum areas as possible, with a level of conceptualisation and difficulty appropriate to the final year of secondary schooling. The AST officer of the OBSSS may appoint an external review panel to advise on the quality and suitability of items.

1.1.4. If requested the Contractor should identify approximately 15 questions out of the 80 multiple choice test questions, each year, which will be omitted by students identified as being from a culturally and linguistically diverse background. These questions will be placed at the end of the 80-question test. Units comprising these questions should consist of the questions in the test that such students would find the most

challenging, in terms of language. These questions, where possible, should represent both verbal and quantitative thinking.

1.1.5. Colour should be used, as appropriate.

1.1.6. Students are permitted to use a paper dictionary, but not a thesaurus. This should be taken into account when questions are constructed.

1.1.7. It is to be assumed that a student's proficiency in written and visual literacy is integral to their achievement in the test. Every effort will be made by the Contractor to ensure students have equal access to the test, but no special provisions (with the exception of the provision in clause 1.1.4 of this Annexure A) will be made to ensure candidates are successful in overcoming specific weaknesses in areas the test is designed to assess.

1.2 CONTENT BY SUBJECT OR DISCIPLINE AREA

1.2.1. Questions will be generic in nature but the context material will be generally taken from Humanities/Arts/Social Sciences and Mathematics/Science/Technology/Business. Material should be drawn from as many domains or curriculum areas as possible with a view to being mapped back to 50 per cent quantitative and 50 per cent verbal thinking.

1.2.2. Care should be taken to avoid the specific content of Years 11 and 12 syllabuses and courses of study in, for example, the selection of themes for particular units or of prose passages on which questions are based.

1.3 CONTENT BY STIMULUS MATERIAL

The range of material considered suitable includes:

1.3.1. Verbal material, involving a variety of modes of discourse (examples in brackets) such as:

(a) narrative (passages from fiction);
(b) description (passages for comprehension and evaluation);
(c) exposition (detailed assertions);
(d) dialogue (comparative points of view); and
(e) instructions.

1.3.2. Tabular material and charts might be drawn from any curriculum area but in particular from:

(a) business practice;
(b) demography;
(c) economics/business studies;
(d) geography;
(e) history;
(f) mathematics/statistics;
(g) politics;
(h) psychology;
(i) science;
(j) sociology; and
(k) technology.

1.3.3. Quantitative material, both with and without a science or social sciences context.

1.3.4. Graphical material from areas similar to 1.3.2 and 1.3.3 of Annexure A.

1.3.5. Visual material, such as maps, diagrams, advertising material, cartoons, photographs and reproduction of works of art.

1.4 CONTENT BY ITEM

1.4.1. Any one unit may contain questions that could touch a wide variety of subject or discipline content. In addition, there should be a wide variety of skills or abilities called upon. In arriving at an answer, the candidate may be expected to work on the given stimulus material in many of the following ways:

(a) exercise verbal skills, such as comprehension of facts, literal meanings, inferences, and causal and other relationships;
(b) exercise short-term recall skills, using memory to sort and interpret information, or to perceive relationships;
(c) interpret data, from verbal and quantitative (including graphical, tabular and statistical) sources, particularly the perception of relationships between different pieces of stimulus material and between parts of a single piece;
(d) evaluate data, where items testing comprehension also call for an estimation of relevance, worth and appropriateness;
(e) exercise reasoning powers; to perceive the logic of relationships and their functions, quality and relevance, such as:
    (i) spot patterns and make generalisations;
    (ii) distinguish major and minor factors;
    (iii) recognise assumptions, both implicit and explicit;
    (iv) predict and evaluate logical conclusions; and
    (v) assess the validity of analogies.
(f) demonstrate judgmental skills, such as the ability to make approximations, assess probabilities, exercise common sense, solve problems, develop practical applications; and
(g) demonstrate visual skills, including comprehension of pattern, spatial relationships, and analogies with verbal and numerical data, symbolic representations.

1.5 CONSTRUCT

1.5.1. The Multiple Choice Test should reflect the skills and processes required for success in university education and allow students to demonstrate a capacity to undertake the intellectual pursuits common to scholarship at senior secondary and post-secondary levels.

1.5.2. Items are to be grouped into two sub-tests, and reported as ‘verbal’ and ‘quantitative’ scores. These two scores will contribute to a total score for the test. The number of items in these two (2) sub-scales must be 40 each.

---

2. SHORT RESPONSE TEST SPECIFICATIONS

2.1 GENERAL DESCRIPTION

2.1.1. The Short Response component of the AST is to consist of a number of questions (approximately 19 – 23 questions) in a 105-minute paper. The number of questions may vary from year to year but there should be sufficient time available for students to be able to think carefully about their responses. It is intended for administration to Year 12 students in the ACT and affiliated overseas schools. Students sitting this test are required to make an original and substantial response to at least half of the questions in the test.

2.1.2. Questions will be generic in nature and based on types of thinking necessary for success in university education. The questions should differ in texture and style from the Multiple Choice questions.

2.1.3. The questions should be designed to measure, in the main, the skills and processes of:

(a) comprehension and interpretation;
(b) analysis and synthesis of information;
(c) critical thinking;
(d) decision-making;

(e) problem solving ;
(f) comparing and contrasting; and
(g) evaluation and response.

There should be opportunities for students to demonstrate lateral and creative thinking.

2.1.4. Any specific information required to answer the questions should be contained in the visual or written material provided for the questions. This information should be pitched at a level of written and visual literacy assumed to be common to students in Year 12. The AST officer of the OBSSS may appoint an external review panel to advise on quality and suitability of the questions.

2.1.6. It is to be assumed the student's proficiency in written and visual literacy is integral to their achievement in the test. Every effort will be made to ensure students have equal access to the test, but no special provisions will be made to ensure candidates are successful in overcoming specific weaknesses in areas the test is designed to assess.

2.1.8. Students are permitted to use a paper dictionary, but not a thesaurus. Calculators are not permitted. This should be taken into account when questions are constructed.

2.2 CONTENT BY DOMAINS OR CURRICULUM AREA

2.2.1. Questions will be generic in nature but the context material will be generally taken from Humanities/Arts/Social Sciences and Mathematics/Science/Technology/Business in approximately equal proportions. Material should be drawn from as many domains or curriculum areas as possible.

2.2.2. Questions should aim at ascertaining what a student can do, rather than addressing specific content of the Year 11 and 12 syllabuses and courses of study, for example, in the selection of the themes or specific information given for the background of the questions.

2.3 CONTENT BY BACKGROUND OF THE QUESTIONS

The range of material considered suitable as background for the Short Response Item questions includes:

2.3.1. Verbal material, involving a variety of modes of discourse (examples in brackets) such as:

(a) narrative passages (passages from fiction);
(b) descriptive passages (passages for comprehension and
(c) interpretation);
(d) exposition (detailed assertions for critical analysis);
(e) dialogue (comparative points of view);
(f) discussion and analysis of visual art; and
(g) instructions.

2.3.2. Tabular material might be drawn from any curriculum area but in particular from:

(a) business practice;
(b) demography;
(c) economics;
(d) geography;
(e) history;
(f) mathematics/statistics;
(g) politics;
(h) psychology;
(i) science;
(j) sociology; and
(k) technology.

2.3.3. Quantitative material in a variety of real life contexts

2.3.4. Graphical material from areas similar to 1.3.2 and 1.3.3 of Annexure B. ACT Scaling Test Contract No: 2010.12909.210 Australian Capital Territory and Australian Council for Educational Research Page 29 of 39 Draft V3 11/11/2010

2.3.5. Visual material, with or without supporting text, such as maps, diagrams, photographs, advertising material, cartoons and reproductions of works of art.

2.4 CONTENT BY ITEM

2.4.1. The Short Response Test should complement and balance the material presented in the Multiple Choice Test so that there should be a wide variety of skills and abilities called upon. Any question could touch upon a wide variety of subject or discipline areas. In arriving at the answer for each question, the candidate may be expected to work in one or more of the following ways:

(a) explain a point of view or make a response to a stated point of view using a variety of skills to support and justify that position;
(b) compare and contrast;
(c) exercise cognitive skills, such as comprehension of facts, literal meanings, connotative meanings, inferences, causal and other relationships;
(d) exercise short-term recall skills, using memory to sort and interpret information;
(e) interpret data, from verbal and quantitative sources (including graphical, tabular and statistical sources, particularly the perception of relationships between different pieces of material or between text and visual images;
(f) evaluate data, where items testing comprehension also call for estimation of relevance, worth and appropriateness; make predictions;
(g) apply information which has either been given or deduced in diverse contexts;
(h) analyse and synthesise information in different formats;
(i) critically discuss points of view in relation to world events and issues and/or to the student's own life and beliefs;
(j) perceive the influences which shape and impede one's own understanding;
(k) find value in unusual points of view;
(l) exercise reasoning powers; to perceive the logic of relationships and their functions, quality and relevance, such as:
    (i) recognition of strength and weakness of argument;
    (ii) validity of assumptions; and
    (iii)major and minor factors.
(m) demonstrate judgmental skills, such as the ability to make realistic approximations, assess probabilities, exercise common sense, solve problems, develop practical applications; and
(n) demonstrate creative skills, including unusual approaches to solving problems or making decisions.

2.5 CONSTRUCT

2.5.1. The test should reflect the skills and processes required for success in university education. It is intended that teachers would be able to apply these types of thinking to the material they teach, regardless of subject area.

2.5.2. The number of items in the Short Response Test should be suitable for the amount of time given (105 minutes). The time available to students to complete this test should include thinking time.

2.5.3. If items are unequally weighted, an indication of their relative weight should appear on the paper for each question.

2.5.4. There should be an indication, on the paper, of the length of response required, either by a statement or by the number of lines left for the answer.

3. WRITING TASK SPECIFICATIONS

3.1 GENERAL DESCRIPTION

3.1.1. The AST Writing Task is a two and a half-hour test requiring candidates to prepare one piece of argumentative writing, of approximately 600 words. It is intended for administration to Year 12 students in the ACT and affiliated overseas schools'. Students sitting this test are required to make an original and substantial response of at least 200 words.

3.1.2. In each edition, stimulus material expressing views about an issue and using a variety of forms or modes of discourse is presented to candidates. Each candidate is required to select a major theme or argument raised in the material, and to develop a point of view about that theme.

3.1.3. Each paper comprises a set of material representing one theme, though the richness of the stimulus material should be such as to allow considerable variety of response.

3.1.4. Students are permitted to use a paper dictionary, but not a thesaurus.

3.1.5. It is to be assumed that the students' proficiency in written and visual literacy is integral to their achievement in the test. Every effort will be made to ensure students have equal access to the test, but no special criteria will ensure candidates are successful in overcoming specific weaknesses in areas the test is designed to assess.[1]

3.2 CONSTRUCT

3.2.1. The AST Writing Task is an assessment of verbal reasoning and writing ability. This task will be used as part of the Scaling Test on the assumption that student ability is validly measured using a component of writing ability in the set of measures applied.

The task aims to assess the ability of the candidates to:

(a) react to stimulus material and focus on an issue;
(b) develop a sustained argument or point of view;
(c) structure and organise ideas in a response; and
(d) control language and form so as to write with clarity, precision and appropriateness.

3.3 CONTENT

3.3.1. Content should be drawn from a wide range of social and/or political issues. At least four pieces of stimulus material should be presented to exemplify the theme.

3.3.2. The mode of discourse, which the candidates are expected to use in their responses, is the argumentative mode. The rationale for this rests on the assumptions that this sort of writing is:

(a) valued by secondary and tertiary educators alike;
(b) consistent with the experience of students;
(c) a good indicator of students' writing ability; and
(d) in some measure an indication of their aptitude for university study.

3.3.3. The stimulus material presented should be accessible to all candidates irrespective of their backgrounds and prior knowledge. The diversity of the candidature should be noted.

3.3.4. The themes embodied within the stimulus material should support candidates by encouraging them to respond to an issue, irrespective of the writing abilities of the candidates.

3.3.5. At least one of the items in the stimulus material should be written in argumentative style.

3.3.6. At least one of the items in the stimulus material should be visual. (e.g. a cartoon, diagram, advertisement, illustration or photograph).

3.4 CRITERIA FOR ASSESSMENT

3.4.1. The criteria for marking the candidates' responses should include:

(a) choice of a major issue directly related to the stimulus material;
(b) complexity of ideas and depth of thinking;
(c) sustained and logical development of ideas;
(d) coherence of an argument;
(e) quality of the evidence of supporting arguments;
(f) distinctive and convincing tone or voice; and
(g) control of language and form including clarity, precision and appropriateness for an intelligent adult audience.

3.4.2. In assessing written expression markers should consider the following issues:

(a) thought and content;
(b) structure and organisation; and
(c) expression, style and mechanics.

Appendix 6

Sections of the Student Information Bulletin for the GAT

WHAT IS THE GENERAL ACHIEVEMENT TEST?

The General Achievement Test (GAT) is a test of your general knowledge and skills in the following areas:

- written communication
- mathematics, science and technology
- humanities, the arts and social sciences.

These areas are very broad. Each represents a body of general knowledge and skills you are likely to have built up through your school years. Because it's a general test, you don't need to do any special study for the GAT. You've already done the preparation in your past study of subjects like English, mathematics, science and history, where you will have built up general knowledge and skills in writing, numeracy and reasoning. These are the knowledge and skills that will be tested. You can get a good idea of the questions by looking at the sample questions and previous GAT papers. Sample questions, previous GAT papers and answers to multiple choice questions are available on the VCAA website (www.vcaa.vic.edu.au). If you try some of these questions you'll see that they give you all the information you need to work out the right answer. That's why you don't need to study for the GAT. You just have to apply your general knowledge and skills.

WHY DO STUDENTS HAVE TO DO THE GAT?

The GAT is an essential part of the VCE assessment procedures. Although your GAT result does not count directly towards your VCE results, it does play an important role in checking that your school assessments and examinations have been accurately assessed, so it is important that you sit the GAT and do as well as you can on all parts of it. This checking can be done because achievement on the GAT is a good predictor of achievement on other assessments. If students have done well on the GAT, then their achievements are likely to be high on their school assessments and examinations as well. Clearly, some GAT questions relate more closely to achievement in particular studies. The VCAA takes this into account when it calculates students' expected achievements in each study for each school. For example, GAT results in mathematics, science and technology play only a minor part in calculating students' expected achievements in humanities studies.

HOW IS THE GAT USED?

The next four sections explain how the GAT is used, why it is important that you do the GAT and why you should do the best you can on the whole test to ensure that you receive full advantage of the VCAA's additional 'safety checks' on your VCE assessments.

The VCAA will use your GAT scores as a basis for:

- contributing to statistical moderation of School-assessed Coursework
- reviewing school assessments in School-assessed Tasks
- special provision and derived examination scores
- checking the accuracy of exam marking.

COURSEWORK AND THE GAT

The VCAA will apply statistical moderation procedures to schools' coursework assessments to ensure that they are comparable across the State and are fair to all students. The statistical moderation process compares the level and spread of each school's assessments of its students in each study with the level and spread of the same students' scores in the external examinations, and adjusts the school scores if necessary. In some studies, statistical moderation uses students' GAT scores as well as their examination scores. This is done where it gives a better match with schools' coursework assessments throughout the State. The examination scores will always have the major influence in the statistical moderation calculations. More information about statistical moderation can be found in the pamphlet 'Statistical moderation of VCE coursework' which is available on the VCAA website.

SCHOOL-ASSESSED TASKS AND THE GAT

The GAT will be used to check each school's assessments for School-assessed Tasks in Art, Design and Technology, Food and Technology, Media, Studio Arts, Systems and Technology, and Visual Communication and Design. If your school's assessments for a particular School-assessed Task are significantly higher or lower than expected from the students' GAT scores, the VCAA will review the assessment of those School-assessed Tasks by sending reviewers to the school to assess the work.

SPECIAL PROVISION AND THE GAT

A student's GAT scores will be included in the calculation of the Derived Examination Score. A student whose performance on an examination is affected by illness, accident or personal trauma may apply for a Derived Examination Score. Additional information on the Derived Examination Score is available in a brochure for students distributed through schools in March and on the VCAA website: www.vcaa.vic.edu.au

EXAMINATIONS AND THE GAT

The examination marking process is rigorous, carefully and expertly conducted and is designed to be fair to all students. For most studies, your examination papers are marked twice, by two different markers. Each marking is done separately and each marker does not know the marks given by the other marker. If there is insufficient agreement between their marks, the paper is assessed by a third marker.

For Mathematics, Accounting, Biology, Chemistry and Physics, examination papers are marked by a single marker. This is because in these examinations experience shows there is little or no difference in the marks given by pairs of markers marking the same paper. In these examinations, the GAT is used to identify which students will have their papers marked a second time. Your GAT score is used to calculate your predicted examination score. If your examination score is significantly different from this predicted score, your paper will be marked again. The GAT is also used as part of a final check on examination scores. If your final examination score is significantly different from the score predicted by the GAT, school indicative grades, other examination scores or School-assessed Tasks for the study, your examination will be assessed again by the Chief Assessor. Your score may go up or stay the same, but it will not go down as a result of this final check.

THE STRUCTURE OF THE GAT

The Two Writing Tasks

Writing Task 1 will present written and graphical information in colour. Your task will be to write a piece that explains the main ideas in this material. Remember that your writing is being assessed, not the extra knowledge you may have about the material. Your writing will be judged on:

- how well you organise and present your understanding of the material
- how clearly you express yourself.

Writing Task 2 will present some views on an issue. Your task will be to express a point of view on the issue presented in the material. You are free to include other knowledge or information that you may have to support your point of view. In doing this task you should aim to present reasons and arguments to support your point of view and to rebut opposing ideas. You must also aim to communicate clearly and effectively to the reader. Your piece of writing will be judged on:

- the quality of what you say about the issue, regardless of the point of view you express
- how effectively you express yourself.

The Multiple-Choice Questions

The multiple-choice questions will cover mathematics, science, technology, humanities, the arts, and social sciences. There will be 70 questions in this section, which will take about two hours to complete. You should attempt every question. Marks will not be taken off for incorrect answers. You must use a pencil on the answer sheet for multiple-choice questions. The instructions will show you how to shade the boxes to indicate your answers. The multiple-choice section will begin with some single questions, with the rest divided into groups of questions or units. Each unit will offer one or more pieces of information and a number of questions about that information.

The following time allocations are recommended.

- Writing Task 1 30 minutes
- Writing Task 2 30 minutes
- Multiple-choice questions 2 hours
- TOTAL 3 hours

http://www.vcaa.vic.edu.au/vce/exams/gat/aboutgat.html

Appendix 7

Summary of comments to two student survey questions about the AST

1. Which questions or parts of the test did you find easiest and why?

This question prompted quite different answers. Some responses were concerned with the subject relationship of the topics.

Some 15 per cent of responses mentioned finding the maths items easy

- *Maths because it has exact answer*
- *Maths questions because they are logical*
- *I just like maths and when there's a single answer to be found, so I don't like most of the short response questions*

Some 5 per cent identified 'English-related' questions as easy.

- *English related questions, for you don't need special knowledge of how to respond, like you do with maths*
- *Arts and English which have more interpreting*

Other responses focused on the kind of item.

About 35 per cent of candidates mentioned that the multiple-choice question were easy.

- *you can work out the answer*
- *the answer is there if you can find it*
- *multiple-choice question the logical process you go through to get the right answer seemed to appeal to my problem solving skills*

About 15 per cent found the short-answer question easy.

- *because you get to justify your answer*
- *I can explain myself*
- *the points you make are more important than the how you write*

About 22 per cent found the writing test easiest.

- *because there isn't a set question you have to answer*
- *I could express my opinion*
- *essay – minimal restrictions*
- *because I can decide the content*
- *more freedom to express yourself*
- *because you can choose your own topic and feel comfortable with it*

- *we had time in WE and could explain our ideas*
- *I love the AST writing task*

2. Which questions or parts of the test did you find most difficult and why?

About 12 per cent mentioned the multiple-choice question as difficult.
- *no room for interpretation*
- *not enough time*
- *way too much maths and science*

About 15 per cent mentioned the short-answer question as difficult.
- *no correct answer*
- *were incredibly vague*
- *no specific question to be answered*
- *didn't know what the right answer was*
- *a page to write a no real directions*
- *difficult to know what is wanted in the short-answer question*
- *difficult to know what the markers are looking for*
- *the open ended nature of the short response was difficult*
- *it is hard to know if you are right or wrong*
- *I object to opinion based questions*
- *undefined short-answer question questions weren't really fair*
- *too much interpretation and not enough logic*
- *hidden meanings and symbolism in the pictures are hard to analyse*

About 16 per cent mentioned the WT as difficult.
- *WT because I am terrible at writing*
- *the open ended nature of the essay was difficult*
- *WT topic out of left field*

Interpretive questions were mentioned as difficult by some.
- *analysing pictures!*
- *finding the underlying message in images*
- *sharks in the lawn was pretty confusing*
- *I struggled a lot with the CD cover*
- *I just don't know how to write a whole page on an album cover from the 1980s*

About 30 per cent of responses mentioned maths as difficult.
- *multiple-choice question and short-answer question had a lot of maths compared with other years*
- *short-answer question biased against people who don't like maths*
- *did not have a calculator I couldn't remember how to multiply or divide*
- *couldn't do maths without a calculator*
- *I need a calculator*

About 10 per cent mentioned English as a source of difficulty.
- *interpretation, I'm not that great at English*
- *compare quotes, not just one right answer*

Appendix 8

The construction of CSSE papers

From Testing to Teaching in the Humanities - Ideas and Resources for Use at the Upper Secondary School Level Drawn from the Commonwealth Secondary Scholarships Examination 1971

Audrey Grant and Graeme Withers

The preparation of test papers such as those used in the CSSE is not suited to individual examiners working in isolation. All resource material for the 1971 Humanities paper and all suggested questions and alternative answers were subjected to prolonged panel discussions. Each panel consisted of ACER staff members augmented by outside specialists, chosen not for their technical expertise in educational testing, but rather for their interest in the relevant subject areas and in teaching at the appropriate levels. The expression of divergent viewpoints and the searching appraisal of the test material by the members of a panel have a direct bearing upon the quality of the final paper. This panelling process brings a range of viewpoints to the preparation of the tests (from the draft stage through to the editing of the final paper), which test-writers or examiners working alone could not. Ambiguities, unwarranted assumptions, and value judgments in the questions which are likely to result from any one person's approach may thereby be recognized in the course of discussion and removed.

After being discussed by the panel, units were edited by ACER staff and prepared for preliminary testing. In this context, it is of interest to consider that if a final test of seven units and 95 questions with a balanced selection of material is required, nearly three times that number of units and questions need to be tried in the preliminary testing.

Students, who were at a level comparable to Australian Grade 10 students, were used as the subjects for the trial testing. Data thus obtained included an estimate of the difficulty (or facility) and a discrimination index for each question, and a reliability figure for the test. The discrimination index provides a measure of the ability of the question to discriminate between 'good' and 'poor' students (i.e., between high scorers and low scorers using the students' performance on the trial test as a whole, or on a particular unit or set of units, as the criterion). A question which is working well will have a correct answer which is significantly more attractive to high than to low scorers, and incorrect alternatives which discriminate in the opposite direction, that is they are preferred by a significantly greater proportion of the lower scorers. Thus each alternative is contributing positively to the task of separating the better student on the test from the weaker student. The great majority of questions included in the final paper discriminated significantly in this way. Questions, which on trial testing had a very high level of difficulty and low discrimination, were carefully examined for ambiguities and either rejected as unsuitable or rewritten. At no time, however, were trial test statistical data the sole criteria for accepting or rejecting units. Rather, these data were used as guides for improving the raw material.

One last panel discussion intervened between the selection of material for a final form of a test and its submission for publication. This panel included senior officers directly involved in the project, together with a number of specialists in the areas covered by each paper who had no previous acquaintance with the material selected. Questions on which there was solid disagreement at the last panel were dropped. It will be evident that by this stage in the test preparation, the answers which are to be scored as correct do not just reflect the test-writer's opinion, but have been scrutinized on several occasions by others with varying viewpoints - and accepted as the best of the alternatives. In sum, every effort was made to obtain as active an interaction between test writers, subject specialists, and other educators as possible. This emphasis on interaction by educators of widely different interests has stemmed from the belief that the educational relevance and quality of the material used in the tests is as important as the statistical validity and reliability of the tests

The authors

Dr Doug McCurry is a Senior Research Fellow at the Australian Council *for* Educational Research. Prior to joining ACER in 1988, Doug was a teacher of English and humanities in secondary schools and the Chief Examiner for Literature in Victoria. Doug has wide experience in the assessment of verbal, humanities and writing tests, and in the training and management of test development teams. He has worked on the development and analysis of multiple-choice, short answer and essay tests. Doug has a particular interest in the assessment of writing in high stakes tests. He designed and managed the writing assessment for the Special Tertiary Admissions Test (1998 to the present), the Graduate Medical Entry Test (1992 to the present) and the Selective High Schools Test. He has directed the ACT Writing Test and the writing component of Cooperative Scholarship Testing Program since 1988.

John Harding is a Research Fellow at the Australian Council for Educational Research. Prior to joining ACER in 2002, John held senior administrative and curriculum positions in independent and government secondary schools in Victoria. He was a teacher of Physics, Mathematics and Science, a member of the VCE Physics assessment group for the VCAA, and is an author of some twenty Physics and Science texts, and associated student and teacher resource materials. John has wide experience in the assessment of mathematics, science and technology related generic and cross-curricular skills. He presently works on the Victorian General Achievement Test, the ACT Scaling Test including the innovative cross-curricular Short Response Section of that test, the Undergraduate Medical Admission Test and the Graduate Medical Schools Admission Test in Australia and the United Kingdom.

ISBN 978-1-74286-401-3